大数据在中国

个人和企业的转型与变革！

赵伟 著



大数据给了中国企业与个人哪些新机遇？
一场关于思维、工作与生活上自我变革！

《给你一个团队，你能怎么管？》作者突破之作！

也许是迄今为止最易懂、最实用的大数据类图书！

江苏文艺出版社
JIANGSU LITERATURE AND ART PUBLISHING HOUSE

大数据在中国

BIG DATA IN CHINA

大数据在中国

大数据在中国

赵伟 著

BIG DATA IN CHINA

江苏文艺出版社
JIANGSU LITERATURE AND ART PUBLISHING HOUSE

图书在版编目(CIP)数据
大数据在中国 / 赵伟著. — 南京 : 江苏文艺出版社, 2014
ISBN 978-7-5399-7375-3
Ⅰ. ①大… Ⅱ. ①赵… Ⅲ. ①互联网络－应用－企业
管理－研究－中国 Ⅳ. ①F279.23-39
中国版本图书馆CIP数据核字(2014)第088379号

书名 大数据在中国
著者 赵伟
责任编辑 孙金荣
策划编辑 一航
文字校对 郭慧红
封面设计 罗久才
出版发行 凤凰出版传媒股份有限公司
江苏文艺出版社
出版社地址 南京市中央路165号，邮编：210009
出版社网址 http://www.jswenyi.com
经销 凤凰出版传媒股份有限公司
印刷 北京兆成印刷有限责任公司
开本 700毫米×1000毫米 1/16
印张 19.5
字数 209千字
版次 2014年6月第1版 2014年6月第1次印刷
标准书号 ISBN 978-7-5399-7375-3
定价 35.00元

目录

导读 大数据正在改变中国

大数据给了中国企业与个人哪些新机遇？一场关于思维、工作与生活的自我变革！本书为《给你一个团队，你能怎么管？》作者的突破之作！

这也许是迄今为止最易懂、最实用的大数据类图书！因为除了本书，再没有另外一本书能让你如此接近中国大数据时代的现在与未来。

全球知名咨询公司麦肯锡最先提出“大数据”概念。随后大数据一词越来越多地被提及，它上过《纽约时报》《华尔街日报》的专栏头条，进入过美国白宫官网的新闻。比尔·盖茨、巴菲特、谷歌公司创始人拉里·佩奇、甲骨文公司创始人拉里·埃利森等商业领袖，都第一时间开始关注大数据。

大数据有三大特点：第一是大，海量的数据；第二是快，可以通过公共数据库快速地获取；第三是我们不再热衷于寻找因果关系，而更加关注于相关关系。

大数据决定着企业的未来发展，正如《纽约时报》2012年2月的一篇专栏文章所称，大数据时代已经来临，在商业、经济及其他领域中，决策将日益基于数据和分析作出，而并非基于经验和直觉。

哈佛大学社会学教授加里·金说：“这是一场革命，庞大的数据资源使得各个领域开始了量化进程，无论学术界、商界还是政府，所有领域都将开始这种进程。”

大数据最核心的作用是可以“预见未来”。华尔街用它准确预测出股票的跌涨，投资机构搜集并分析上市企业声明，从中寻找其破产的蛛丝马迹；美国疾病控制和预防中心依据它，分析全球范围内流感等疫病的传播状况；美国总统奥巴马通过它获取更多的政治选票。

如今，大数据已经到了中国，它将改变中国，改变一切传统企业，改变整个市场的格局。我们的工作、生活、社交都将与它息息相关。从移动、电信、阿里巴巴、微博、微信、百度到互联网的每一个角落，整个互联网就相当于一个大数据库。

了解大数据，学会其中的原理与运用的技巧，也许一个百度搜索热词就能给你一个商机，你的想法创意很快就能通过各种公共数据库得到验证。退一步来说，就算你开一个网店，起点也比他人要高出许多。高端论坛中经常会提到的“互联网思维”如果离开大数据的支持，就会如无源之水，无本之木。

本书用最精简的文字讲述、分析了大数据的特点、原理以及在当下中国的各个行业和领域的运用。同时也给我们的企业与个人提供了新的

参考：大数据时代将对中国企业转型提供哪些支持？我们的优势和劣势在哪里？我们该如何搭上大数据的快车，实现商业价值与个人理想呢？

如果说您已经通过许多同类书籍知道了大数据是什么，那么本书则更注重为您解惑和提供实用指导：面对大数据，我们应该怎么做？书中不仅展示了谷歌、微软、亚马逊、Facebook、Twitter等大数据先锋公司最具代表意义的应用案例，也展示了阿里巴巴、腾讯、百度、奇虎360、小米等首批搭上大数据快车的企业的中国式突围。

序 大数据在中国

这本书要解决的第一个问题是，我们应该如何理解“大数据”？

大数据的英文表达是Big Data，意思是“海量数据”。数据的规模大到了已经无法用当前的技术和工具来处理，那就必须突破瓶颈，从而产生数据革命。对数据的处理包括很多方面，有收集、整理、分类、存储、分析、预测和输送等等。

大数据就这么简单吗？当然不是。数据如同人体的血液，大数据则是整个人体系统与血液有关的部分。最早涉及这个概念的是天文学和基因学领域，因为这两个学科非常依赖对数据的分析方法，尤其是对“海量数据”的分析。它也是电脑和互联网结合的产物，因为电脑实现了数据的“数字化”，让它们像数字一样容易储存，互联网则实现了数据的“网络化”，让它们通过网络可以自由快速地传输。

从此，大数据才真正拥有了无穷的生命力。互联网的技术不断发展，渗透进我们的工作和生活，加上移动网络、物联网与其他各种联网设备的出现与普及，一个必然产生的现象就是数据的迅速增长。有90%的数据是互联网出现以后才产生的，它以指数级的速度在我们的生活中不断增加，从海量至于无穷大，世界正被数据淹没。

我们需要更加关注的，是数据从量变开始质变，并且体现在多个方面，触发蝴蝶效应，推动其他领域的变化。

第一，催发数据性思维。

一种全新的思维模式，这首先是大数据时代的到来带给我们的。思维方式的改变，推动产业的根本性变革。这种变革甚至是颠覆性的，因为思维是最大的生产力，也是社会文明进步的决定性力量。

历来的商业变革，其发端都不是某种技术，而是思维方式的转变。思维产生需求，需求推动技术。人们有了新的思考，有了新的视野，才开始低头审视旧的经济体制和传统的商业理念，去思索和创新商业逻辑，研究更实用、先进和高效的模式。反之，如果人们不能与时俱进，吸收并且创造顺应潮流的新思维，再通过新思维去重新组织资源，建立架构和制定策略，那么貌似强大的体量反而会变成继续前进的累赘——就像19世纪的清帝国。

这种通过新思维来颠覆旧秩序的案例，不仅发生在国家的竞争之间，还在信息技术领域频繁上演。它总是最先由信息与技术的新思维起步，然后渗透到传统领域，继而蔓延到全社会，改变每一个行业和每一个人。

相关案例到处都是，比如黑莓、摩托罗拉、诺基亚、柯达、雅虎等

公司，比如华为、联想、中兴、百度、360、中国移动以及它们的竞争对手——谷歌、思科、微软、三星、西门子等。

数据性思维表现在：

1.对全部数据进行分析，而不是随机抽样；

2.并不过于追求精确性，而是重视数据的复杂性；

3.更多挖掘数据的相关性，而不是因果关系。

虽然对于后两条在不同的领域存在争议，在不同的需求面前也存在侧重点的区别，但就数据性思维而言，这三条特征足以让我们的世界天翻地覆。当然，许多超级巨头的没落也并不是因为它们没有数据性思维——原因总是很复杂，但它们归根结底总是倒在自己的思维上。它们被新的思维超越，并且彻底击倒；如果不奋起直追，就会被淘汰，退出历史舞台。

昔日巨人在感慨自己未及早行一步，追赶者在叹息何处寻觅超级思维人才。而在今天，数据性思维就是最新的智慧，虽然它的威力还没有进化到可以导致巨头轰然倒下的地步。不过，需要警告的是，如果你今天不给予它足够的重视，在下一拨没落帝国的名单中，你的名字将会赫然在列，成为失败者中的一员。哪怕你现在风光无限，在被超越和打败的时候，也会无能为力。

第二，产出“数据资产”。

思维充分展示了它的力量，而在大数据时代，另一个变化是数据成为一种资产并被产出。这是因为，当我们需要更加全面的数据来提高预测分析精度时，就必须制造更多的便宜、方便和自动的数据工具，收集、加工或生产数据。就像你住的房间越大，需要的佣人就越多，他们的工作量就越大，而你的需求是佣人最好不犯错误，干活越快越好，越多越好，给你提供完美的生活。

这些工具包括什么呢？上网时我们使用的浏览器、软件、窗口和后门工具，它们会记下你的各种个人信息数据；生活中我们使用的智能手机、GPS导航器、智能手表或其他数码产品，时刻都在产生数据；还有路由器、智能玩具甚至电视机、冰箱等，如果它们联入网络，在为你服务的同时，也在产出无数的数据。

当你和女朋友出去购物时，你需要刷卡、移动上网，那么超市收银的刷卡器、路由器、无线WLAN和3G网络、银行ATM以及监控摄像头，都在收集和产生与你有关的数据。

除非你躺在家里什么都不干——扔掉手机、电脑、固定电话和所有的银行卡，也不出门购物，变成彻头彻尾的隐居者，从这个世界消失，否则你时刻都是一个产生数据的源头，并被接入一个无限大的数据互联

网。全世界几十亿人共同组成了一个庞大的数据生产工厂，提供了几乎无法计量的数据资产。

为什么我说这是一种资产呢？因为数据在互联网领域意味着流量，而流量就代表着财富。流量拿来变现，可以是广告，可以是游戏，也可以应用于电子商务。拥有这些数据的人或企业，就能由此获得巨大的利益。这时，商机就产生了。

因此，在大数据时代开启之前，意识到数据的资产价值的公司就已开始布局。它们早就盯上了数据生产的源头——消费者，通过出售数据工具来把人们接入互联网，获得人们的数据流量。

第三，数据资产可以变现。

虽然在事实上，任何资产都是可以变现的，但没有哪种资产的价值可以超过“数据”。我们拥有了数据资产，就必须通过分析来挖掘它有多少价值，通过预测来确定价值的方向，通过整理来划分价值的类型，最后再“变现”为用户价值、股东价值乃至全社会的价值。

在这个过程中，核心环节就是预测——这也是大数据分析的目的。数据是海量的，也是复杂的；数据无孔不入，无所不至，怎样采取实用技术来进行划分？如何通过数学建模来从数据中整理价值走向，建立价值链？最关键的是，我们要预测事态发生的可能性，然后制订相应计划，采取合理措施。

这方面的应用十分广泛，像流行病、股价、经济走势的预测。当预测可行并成为一种机制时，人们就有了更进一步的可能——通过适当的干预，来引导事情向着期望的方向发展。

例如，针对用户的个人喜好和对消费能力的统计，电商网站向他们推荐不同的商品，引导他们多消费。游戏公司会在游戏运行过程中对玩家的行为数据进行分析，及时调整针对不同玩家的计费点和关卡难度，也对玩家的消费欲望和能力进行分析，最大化地从玩家身上获得收入。现在所有的网络公司都已拥有了“精准营销”的能力，通过恰当的技术手段向不同的用户展现不同的广告（个性化广告），以此来提高广告收入，也帮助自己的广告用户提高产品销量。

这些例子都很好地证明了数据资产的价值，以及价值的可塑性。它也是一个巨大而完善的产业链，它的分析、预测和控制技术，亦广泛地应用于其他行业，对我们个体的生活、工作和思维的提高都有不可估量的帮助。

本书抛出的第二个问题是：大数据非常热，但如何才能把它变成一个广泛实用的利器而不是阶段性的泡沫呢？

我们都有这样的感觉：大数据正在逼近和包围我们，因为现在人人

都在讨论，企业在探讨，政府也在研究，新闻铺天盖地，理论遍地开花。数据也在不停地产生，个人、企业、政府机构、互联网，都在产生数据。各式各样的数据正在冲击我们，颇有“乱花渐欲迷人眼”和“身在庐山”的感觉，难免感到迷茫与困惑。

我们究竟应该从大数据中挖掘到什么？

企业和组织怎样通过大数据完成蜕变和转型？

大数据技术应该怎样实用化、有效化，让它适用于每一个人？

这些是人们普遍关注的问题，要想实现这些具体的目标，我们必须先破除障碍。障碍既是技术上的，又是心态和认知上的。技术障碍最容易攻克，心态和认识上的障碍往往很难在短时间内消除。

从国内这几年大数据发展的现状来看，人们从数据中已经发现了新的价值，在技术层面也已经做足了准备，但之所以大数据市场还未能大规模地启动（除少数几家公司外），原因在于心态、认知的滞后和旧观念的羁绊。

上海一家IT公司的经理朱先生拿他亲身经历的一件事举例。不久前，朱先生开车在路上行驶，被一辆摩托车剐蹭了，由于肇事者拒不认账，朱先生只好请求有关的交管机构调取马路上的监控视频，希望对这场事故责任有准确的认定。这当然是一个好的方法，但他得到的答复是：“你的请求是可以的，不过需要将你的汽车押在交管部门15天，相关数据也要在15天后才能调取出来。”

朱先生很苦恼：“明明一件很容易、很迅速就可办到的事，为何弄成这样？”

这就是观念或心态的问题。我们知道，即便在中国一个较小的县城，马路上和关键的区域也都装上了摄像头；在大的城市比如北、上、广、深，传感器早已经星罗棋布，到处都是。收集和记录视频是一件信手拈来的易事，这些数据也具备了为社会、为民众提供更多优质服务的可行性。同时，现在的大数据技术也已经完全可以高效处理和分析这些数据。可事实往往相反，人们依然无法从中获益，有的机构最在乎的是自己拥有多少权限，而不是提供多少服务。当然，相关的机构也对此有怨言，他们也有自己的理由和无奈之处。原因就在于人们服务意识的缺乏，以及旧的思维和心态的顽固影响，这便构成了一个牢固的阻力场，让每个人深陷其中，都想出来，但又都拔不动腿。

大数据不仅应用空间广泛，也有巨大的需求空间。不论在医疗、交通、航空、电力、电信、金融、城市管理等公众服务领域，还是在制造业、创新行业等各式各样的企业当中，都存在从数据中发现新价值，提供更个性化、更有价值的服务的需求。也就是说，我们已经完全准备好

了技术，但还没有准备好观念和意识。现在中国正大力推进智能电网建设，虽然智能电表安装很快，但后端的数据整合远远跟不上，导致用户虽已装上了智能电表，但得不到应有的服务。比如当电表只剩下最后几十度电时，都无法获得短信提醒服务。这和朱先生的遭遇如出一辙，不是技术上做不到，而是观念需要改变。

大数据方兴未艾，正在源源不断地释放出更多的能量，推动我们的产业升级、社会转型和企业转型。我们要想从中获益更多，就需要全民参与、全民破冰、全民转变观念。只有这样，数据才会成为我们的“新能源”，而非“烫手山芋”。只有首先破除这些障碍，大数据在我们的生活中才真正具有无限广阔的应用空间，成为推动社会发展的核心动力。

要想成功破冰，更新旧的观念，我们必须从内部数据和内部问题切入。只空喊口号、形而上地研究似是而非的问题是起不到多少实际作用的。在不同目标、不同状态下的企业、部门和个人，其实都可以开启自己的大数据之旅，从大数据中挖掘到更多价值，但一定要脚踏实地，拒绝空谈，从实际出发，从小处做起，从细节开始，打下坚实的基础。

只有从现有的数据入手、以拥抱客户为核心任务、以解决内部问题为基本的出发点，才能切实地推动我们的大数据发展，成功开启破冰之旅。比如公交系统，在整合了各种公交平台的信息体系后，将公交车、地铁和出租车的出行进行“三位一体”的融合，人们就可以用手机实时地查询正在等待的公交车还有几分钟到达，确认实时的路况和交通拥堵信息，灵活选择自己的出行工具和路线。

当前，大数据对于企业与机构的显著应用效应主要体现在两个维度：一是在前端发现更多的商机，对客户实现全方位的关照，挖掘和拓展更大的增值空间；二是管控风险、提高效率和降低成本。

有一位在电信公司任职的项目经理对我说：“现在电信之间的竞争日趋白热化，市场也将饱和，怎样快速地掌控市场和用户的动向，制定个性化和针对性的营销方案，就成为在激烈的竞争中获胜的关键。我们在采用新的大数据方案之后，直接提升了服务速度和个性化的服务效率，数据中心的反馈时间从原来的两周缩短到现在的不到两天。这样一来，我们的客户经理就能够实时获取大量信息，更有效地分析客户的使用习惯，占领服务客户需求的先机。”

这就是在解决问题的基础上成长，在挖掘数据价值的前提下破除障碍。对大数据在现实中的具体应用，尤其对企业来说，首先必须以客户为中心，制定战略规划和自己的大数据蓝图。其次是根据自己的业务优先级，建立数据分析体系，逐步提升大数据的分析与应用能力。最后就是定制可量化的指标，获得投资回报率。

从目前的大数据实践来看，我们的应用主要体现在对现有数据的整合、挖掘，进而去提升客户端的应用价值和带动企业的效率。如果想让数据革命真正变成席卷全球的第三次工业革命，产生新的生产力，还有很大的创新空间。

对中国而言，一个伟大的时代才刚刚开始。

赵伟
2014年元月10日 北京

CHAPTER 1 大数据，你还不知道的部分

FB数据单元——信息导航图

数据是由什么组成的？

一个数据单元有多大？怎样产生和传送？

这是我们首先要知道的基本问题。曾经有人把数据比喻成花粉，蜜蜂搬运花粉使果实得以产生。每一个花朵都是数据产生源，蜜蜂承担着数据搬运工的工作。我认为这个比喻非常恰当，但有更好的概括——数据就像人体的血红细胞，一个数据单元就是一组营养单元，由肝脏产生，输送到身体各处，供应器官的需要。

数据单元是信息传输的基本单位。特别是在网络中，一般的网络连接不会允许将任意大小的数据包进行传送，它有严格的规则，采用分组技术将一个数据分成若干个很小的数据包，并且给每一个小数据包都加上它的属性。这个属性是与传输有关的，包括源IP地址、目的IP地址、数据的长度等。

和血液一样，它有固定的目的地。所以，我们把一个这样的小数据包称作数据单元，也可以称为数据帧或帧。如此一来，数据信息流的特点就明确了，每次要传送的数据都是特点鲜明的“包裹”，它们的规格和封装方式都是相同的。这有利于数据传输的标准化，也简化了它的产生、加工、包装和传送方式，使得大规模应用数据成为了可能。

我们发现，任何一个数据组织都有它的既定体系。在这个体系中，可以划分为位、字符、数据元、记录、文件和数据库六个层级。前一个层级的数据元组合产生了后一个层级，最终实现了更大规模的数据集合。

在这六个层级中，“位”数据处于第一层，一般的用户不需要探究，但后面五个层级则需要我们掌握，因为它们是人们在输入和请求数据时要应用到的。

当不同的数据包或数据元素之间存在着特定关系（一种或很多种）时，它们就构成了数据结构，也就产生了“电脑存储和组织数据”的特定方式。人们认真选择的数据结构能够带来更高的运行或者存储效率。这时，检索和索引技术的需求就随之产生了。更好的技术可以让我们的检索更加高效。

我的朋友沙尼尔是一位任职于谷歌公司的大数据专家，他在去年出版的名为《数据算法与应用》的书中对于数据的性质这样解释：

“数据结构代表着一种联系，它是数据对象及存在于该对象的实例和构成该实例的数据元素之间的各种联系。同时，这些联系可以通过定义有关的函数给出并量化。”

数据对象又是什么呢？沙尼尔认为，一个数据对象是实例或者值的集合，而数据结构是抽象数据类型（ADT）的物理实现。他将一个数据结构的设计过程分成抽象层、数据结构层和实现层这三个层级。在这其中，抽象层是指抽象数据的类型层，它讨论的是数据的逻辑结构及其运算，数据结构层和实现层则更贴近于形象化和实用性，它们讨论的是一个数据结构的表示和在电脑中的存储细节以及这种运算的实现。

如果我们结合现实应用，将数据结构解剖开来，会看到什么？你立刻就会发现自己已经漂浮在数据王国的海洋之上，它们离你是如此之近，并时时刻刻与你的生活发生着关系。

● 字符

当我们输入一个字符时（通过键盘或其他设备），系统会直接将字符译成某特定的编码系统中的一串位的组合。一个字符在电脑中占8位，即一个字节。这就是字符，也是一般而言数据的最基本单位。同时，电脑系统可以使用不只一种编码体制来处理字符。比如，某些系统将ASCII编码体制用于数据通信，而把EBCDIC编码体制用于数据的存储。广义上，我们在纸上写下一个汉字单词、一个阿拉伯数字，也可视作“数据”中的一个字符。

↓

● 数据元

数据元是数据的层次体系中最低一层的逻辑单位。我们为了形成一个逻辑单位，需要将若干位和若干的字节（字符）组合在一起。比如一句完整的话，一段完整的富有逻辑的代码，一个最小的信息流等。因此，数据元也可称作字段。它是泛指的，其中的数据项才是数据实体，比如一个完整的手机号是一个数据元，138或后面的数字按段分开，则是具有单独存在意义的数据项。

↓

● 记录

数据元以逻辑相关的形式组合在一起，就形成了一个数据记录。价值在这时候开始陡然提升。比如一条员工记录——编号、姓名、性别、职称、所属部门——包含了若干的数据元，它们之间有逻辑相关性，再加上辅助性的数据项，就构成了完整的记录。这是数据库中存取的最低一层的逻辑单位。

↓

● 文件

一个完整的文件是由信息和介质构成的，它是被命名的、存储在某种介质上的一组信息的集合体。比如一篇文章、一张唱片、一份合同，

甚至于一本书，都可称为数据元件。一个文件在逻辑上可划分成若干的记录，那么文件就以记录序列的形式体现。文件与存储介质无关，介质的改变不会改变文件的性质和它的价值。

↓

● 数据库

数据库是最大的层级，它是一组有序数据的集合。在这组有序数据中，包含大量的文件——这些文件之间互相又具有逻辑相关性，并以某种检索价值被标注。根据不同的应用需求和不同的领域，人们有时也将数据库分成若干段，而不是唯一存在。数据库有备份，可以随时检索、整理和利用，也可以随时被有权限的人更改。

核心：整理、分析、预测、控制

“大数据”的核心并不是我们拥有了多少数据，而是我们拿数据去做了什么。如果只是堆积在某个地方，数据是毫无用处的。它的价值在于“使用性”，而不是数量和存储的地方。任何一种对数据的收集都与它最后的功能有关。如果不能体现出数据的功能，大数据的所有环节都是低效的，也是没有生命力的。

☆整理

整理有两个目的，一是将所有的数据归类，把它们放到该去的地方；二是利于我们检索，随时调取数据进行利用。这和我们整理书架的目的是一样的。面对同样的数据，不同的整理方法决定着我们的效果是好还是坏。

美国国会图书馆的检索工程更新很能说明“整理”的重要性。在国会图书馆，人们曾经经历过一段困难时期，因为信息量随着网络技术的发达不断暴涨，就连保存的推特（Twitter）信息（只是图书馆数据中很小的一部分）就达到了接近两千亿条，存储文件的体积更达到133TB。删除是不可能的，因为每一条信息都已经在这套社交网络中获得了读者的分享与转载——那么，如此庞大的数据应该如何整理？

技术团队需要想尽一切办法、穷尽所有智慧才能拿出切实可行的检索方案，让图书馆的用户可以方便地利用这些信息。也就是说，技术人员必须着手建立一套帮助研究人员（包括其他用户）快速访问社交平台数据的系统，因为随着网络工具和文化潮流的不断发展，人们都在趋向于电子阅读而不是来看纸质书。

从2000年开始，图书馆就启动了整理归档的工作——那时的难度较小，因为尚未接入社交网站，政府内部的系统储存的数据在一定时间内是静态的，增长速度较慢。虽然数据的总量也超过了300TB，但工作人员觉得：“总有一天可以整理清楚。”

然而，推特的出现令图书馆的归档工作陷入了痛苦的僵局。图书馆方面实在找不到合适的办法来保证信息易于搜索，在这个过程中还不能出现无法容忍的错误。如果继续使用旧的方式——磁带存储，那么仅查询一条2006年到2010年之间的推特信息可能就要耗费一天，如果查询期限再加上一年，所要的时间就要增加四倍。

国会图书馆的一位工作人员费舍尔说：“我们在庞大的数据面前感到头疼，整理成为了一个不可能完成的工作。如果无法把它们归类，这些数据就变成了包袱，需要它们的人检索不到，我们却又不得不保管它们。”

推特的信息之所以难于整理，一方面是由于它的数据量过于庞大，另一方面的原因则十分现实，因为每天都会有新数据不断地加入进来。就像我们的微博一样，每分钟都有大量的新信息产生，人们不断在发微博。所以，这种增长速度会不断地提升，要用传统方法把它整理好，几乎是不可能的。

此外，这类信息的种类也越来越多样，比如普通的推特信息、利用软件客户端发出的自动回复信息、手动回复信息、包含链接或者图片的数据等等。经常使用微博的人对此心知肚明。传统方法在新的数据更新特点面前，根本无从下手。

费舍尔说："如何寻找解决方案？道路是曲折的。我们开始的时候考虑分布式及并行计算方案，但这两类系统实在太过昂贵。要想真正地实现搜索时间的显著降低，就需要构建起由数百台甚至几千台的服务器构成的庞大的基础设施。天！想想都不可能，这对于我们这种毫无商业收益的机构来说，成本实在太高了，一点也不符合实际。"

图书馆最后找到了大数据工程师。专家针对图书馆的具体情况，给出了一系列的实用方案。开源数据库工具Raik的创始人菲利普斯建议采取分类处理的方式，即利用一款工具处理数据存储、一款工具负责检索工作，另一款则用于回应查询请求，非常简单有效地完成了整理的工作，让海量的新信息与庞大的旧数据完美融合，也保证国会图书馆实现了数据库的更新换代。

在整理完成以后，数据的总量增加了几十倍（每时每刻仍在增加），检索速度反而比以前更快，甚至已经实现了检索结果瞬间到位。

☆分析

分析是指对于数据进行"有效分析"。数据往往规模巨大，成分复杂，且来源不一。尤其在大数据时代，数据往往同时具有四个特点，简称4个V：数据量(Volume)大、速度(Velocity)快、类型(Variety)杂、价值密度(Value)低。怎样在最短的时间内做出最有效的分析，就成了一项核心工作。

随着大数据时代的来临，大数据分析也紧跟着应运而生。而且，传统的数据分析也在与大数据分析进行融合。

目前人们对于数据的解决方法主要还是这几个方向：数据怎么做预处理？归档的文档怎么能够及时查询？如何使用你的挖掘和分析技术来看到视野范围内的全息的大数据内容？在海量数据面前，传统的分析方式是做不到的。

数据分析的弱点也是需要我们警惕和谨慎思考的。去年六月份，有一位投行的华人高管蔡先生找到我。他正在考虑是否要退出欧洲市场，

因为经济形势太不景气了。他觉得将来一定会发生欧元危机，一旦危机爆发，公司就会陷入破产的困境。

没错，经济有可能低迷，这是一个潜在的事实。但是，我提醒蔡先生注意另一个事实，那就是这家投行在欧洲已有近五十年的经营史，树大根深，有了很庞大的市场，也有大量的老用户。假如这时退出欧洲，会不会让人们觉得这家投行一遇到风吹草动就弃械投降、根本不值得信任呢？

蔡先生恍然大悟，他马上决定不能清算公司在欧洲的业务，不管未来有什么危机都要坚持下去，即便在短期内付出巨大的代价，也在所不惜。在做出这个决策时，蔡先生并没有忽视那些经济层面的数据，在我的建议下，他采用了另一种不同的思维方式，在数据的考量中纳入了更多更全面的信息。在困境中做出正确决策的人和机构，往往能够赢得更多的尊敬，而这不是传统的数据分析可以捕捉到的。

蔡先生的故事在告诉我们数据分析的威力之外，也充分体现了数据分析的短处和局限。虽然人类的生活现在由收集数据的电脑在调控指挥，当人的大脑无法及时理解和判断情况时，数据也可以帮我们解读和分析它的意义，并且帮助我们弥补对于直觉、情感的过分依赖，减轻我们内心欲望对于理性的扭曲。但归根结底，数据并不能代替人的思考，只有明确数据的真实价值，才有助于我们摆脱对数据的完全依赖。

真正的大数据分析就是要帮我们搞明白数据的真实价值，它在研究大量数据的过程中寻找模式、相关性和其他的有用信息，来帮助人们和企业更好地适应变化，并且做出那些真正明智的决定。

在大数据的层面上，对海量数据有四个不同的方向和解决工具：

1.技术上解决了廉价数据的问题；

2.几乎可以实时地对数据进行分析，而不会有任何滞后，保证了数据的实效性；

3.大数据的可视化和发现性，使得搜索与可视化成为热门应用，也让数据更加精确；

4.在设备层面，拥有了经过优化的一体机设备，使得数据制造和分析更加便捷，成本也更低。

即便拥有最好的技术，在对数据进行分析前，人们也应该先了解数据的真实含义——就像了解自己一样。如果你对于数据是陌生的，那么作为一个决策者来说，你对于自己的事业就是十分危险的。现在许多产品经理、设计师和高管在没有完全理解数据的真实含义的情况下，就直接根据数据来修改自己的产品设计、做出完全基于数字逻辑的决策，结果往往事与愿违，导致糟糕的结果。

☆预测

大数据技术就像一面细致入微的显微镜，不但能够收集和分析最不起眼的信息，而且能够基于这些信息之间的逻辑关系做出科学决策。就像我们可以根据人的表情与言词判断他接下来的行为、量度他内心的情感状态一样，预测功能在商业、经济乃至其他领域都有助于政府和企业管理者做出更多的理性决定，而不仅仅是依靠直觉和经验。

IBM公司的能源电力应用部门经理布兰德说："我们运用大数据预测风电和太阳能，精确地预测来自太阳能和风能的电力产出，取得了很好的效果。这是一种前所未有的创新模式，将使能源电力行业解决可再生能源的间歇性缺陷。"

IBM公司开发了一种结合天气和电力预测的智能系统，提高了系统的可用性并优化了电网的性能。它是足够改变游戏规则的新发明，结合大数据分析和天气建模技术而成，是现在全世界最先进的能源电力解决方案，可以提高可再生能源的可预测性。

这项名为"HyRef"(混合可再生能源预测)的大数据预测技术，利用天气建模能力、先进的云成像技术和天空摄像头，接近实时地去跟踪云的移动，并且通过传感器来监测风速、温度和方向。通过精确的分析，能为风电企业提供未来30天的区域内的精准天气预测，或者未来15分钟的风力增量。这就使能源公司有条件将更多的可再生能源并入生产线，减少碳排放量，然后制造更多的清洁能源。

这种预测能力让我们的生产模式得到真正的升级，而且可以应用到其他领域，比如天然气、煤炭或其他传统行业。不仅在实体产业，非制造业的服务产业对于大数据预测的需求更盛，也有着更广阔的市场。例如，可以帮助企业和政府机构进行业务（服务）分析与预测，对工作量身定制，降低成本，事先应对危机；再比如，可以对房地产销售的价格走势进行预测，它的精确性远远超过传统的房地产分析师。我们每个人都将从中受益无穷。

☆控制

如果你正确地使用了大数据，收集、整理、分析和进行预测，它将为你提供梦寐以求的情报和洞察力。它的控制功用是如此强大，既能够让你掌握最全面的信息，也足以使你从容引导——使自己免受威胁，保护企业，解决潜在问题，并通过自检和优化提升效率。

现在全世界每天都要产生超过3EB的数据，我们有理由相信，随着互联网、各种移动平台越来越广的拥有率和使用频率，这个数字正在不断升高。从棱镜门事件中我们已经知道，美国政府千方百计要加以运用的就是这些数据——以大数据技术来把它们吸纳进去，除了用于正面

（反恐），也在试图监控和控制民众。

有一家美国的顾问公司预测，在今后，美国国内还需要10多万个数据分析专才，以及100多万名能够运用数据的经理人。由此可见，大数据的应用在美国已经十分普及，他们将大数据大量地运用于社交媒体、移动网络和对舆情的分析上，进而达到控制选民、管理资讯和监控敌国的目的。

谁对大数据的研究越早，准备越充分，谁体现出来的控制力就越强。毫无疑问，美国人已经走在了最前面。

控制的基础是管理好这些大量的非结构数据，假如管理得当，我们就能从中挖掘出有效信息，实现企业和政府的管理革新。有先见之明的公司都正在从内部的各种来源以及云基础设施中收集越来越多的数据，它们构建可自控的数据中心，聘用和培养自己的大数据工程师。但还有更多的企业仍然徘徊在门外。后者注定会让自己的企业远远落后于人，它们没有办法获取及时有效与海量的信息，以及由此产生的洞察力，自然也就做不出明智的决定。

2013年，我们与安全公司EOA北美分公司在东亚地区共同完成了一项大数据调查。调查的对象是300位来自中国各行各业的高级主管。结果发现，已经有49%的中国公司关注或者非常关心大数据管理问题，但还有38%的中国公司并不明白什么是大数据，对大数据还是一头雾水；另外有27%的中国公司表示他们对此并没获知太多信息，只知道细枝末节或停留在看客阶段。

另外，我们还发现76%的中国公司没有使用恰当的工具来管理自身的系统数据（IT系统），而是采取其他的独立或缺乏互联功能的系统。有的公司甚至还在采用电子表格的方法对数据进行记录和管理。

这是一次令人灰心的调查，但可喜的是，我们看到了积极的增长速度。相比于2012年或更早的时间，投身于大数据的中国公司正以疯狂的速度增加。随着设身处地感受到它的好处的公司越来越多，人们已不再准备持观望态度，而是立刻参与进来。

要实现大数据控制的关键之一是“日志管理”，整合与自己有关的所有数据，比如企业日志，建立索引库，然后设计用户易于理解和使用的界面。要把数据充分利用起来，就必须使数据关联化和规范化，具备报告、反馈与防卫入侵的能力。每一家成功的电商网站和面向用户的企业官网，都是这么做的。

现实的情况是，国内目前只有56%的受访者使用日志管理的解决方案来管理他们的数据。很多公司使用电脑系统自带的普通日志或者建立一个电子表格进行这项工作。更有39%的受访者向我们表示，他们根本

没有对日志（数据）进行管理。

“有什么用吗？”他们问。这表明，国内对大数据核心的认识和应用任重而道远。提高认识和加强推广成为了当务之急。

此外，相关的技术更新、方案和平台必须跟上新信息产生的速度。数据的产量以几何级的速度增加，它比宇宙中的星星还要浩瀚。如果我们检索数据的时间太长，分析和预测就失去了意义，控制与管理更无从谈起，还会造成严重问题。

大数据先行者

中国正处于大数据的起步阶段，国外先行者的经验对我们具有十分重要的参考和借鉴价值。全世界的富有远见者早在多年前就已经开始了你追我赶，在自己大数据中心的建设上各显神通，力争在这场战争中取得先发优势。

☆英特尔（Intel）

英特尔公司是全球最大的半导体芯片制造商，成立于1968年，具有几十年产品创新和市场领导的历史。全球第一个微处理器就由它在1971年推出，从而引发了计算机和互联网革命。从硬件入手以配备大数据需求是英特尔首先做的准备，同时对于软件也毫不放松，在Hadoop系统、Hbase、HDFS上都做了增强和优化，并且推出了Intel Hadoop Manager 2.0。

2012年7月，英特尔对外发布了自己的Hadoop商业发行版(Apache Hadoop Distribution)，成为几家大型厂商中唯一拥有自身发行版Hadoop的公司。

☆IBM

IBM以对数据挖掘和数据分析领域的收购展开了大数据时代的布局，后来正式推出名为“3A5步”的动态路线，然后结合信息管理、业务分析等软件提出了属于IBM的大数据平台架构。

该公司的大数据架构涵盖了IBM在大数据领域的四大核心能力和相应的产品线，包括：Hadoop领域的InfoSphere BigInsights，流计算领域的InfoSphere Streams，数据仓库方面的InfoSphere Warehouse和etezza以及信息整合与治理(Information Integration and Governance)方面的产品Optim及Guardium。

☆Hortonworks

2011年从雅虎剥离后，Hortonworks公司在当年8月份就发布了一款基于Hadoop的数据平台的技术预览版(Hortonworks Data Platform，HDP)。仅过几周，该公司又推出了基于Hadoop 0.23的HDP 2.0版本，该版本的Hadoop获得极大提升，实现了下一代的MapReduce。

尽管成立时间很短，但Hortonworks行动迅速，就在IBM宣布了基于Hadoop的大数据分析平台后不久，它便开启了自己的大数据战略。此外，它还与Talend公司达成协议，将在其数据平台上提供给Talend公司Open Studio for Big Data工具，以全面应对大数据处理。

☆微软（Microsoft Corporation）

微软公司作为传统的IT业旗帜企业、当之无愧的垄断巨头，进入大

数据领域看起来却并不是第一位的。它经常被人们认为起步较晚，但其实微软早在2006年就致力于研究类似Hadoop的开发计划Dryad，并使其获得产品化。微软一直保持自己的独特风格，不紧不慢，但从不在关键领域落后于人。

2011年初，微软公司发布了自己的并行数据仓库项目（SQL）。一年后，正式发布了SQL Server 2012数据库平台，把业务延伸到了非结构化数据领域。当Windows Azure Marketplace和SharePoint等工具推出以后，微软公司厚积薄发，已完全具备了打造大数据平台的能力。

☆思爱普（SAP）

成立于1972年的思爱普公司在软件领域一向具有极大的优势，而且其产品大多聚焦在对数据的分析能力上。这使它在大数据时代开启的一瞬间，就已处在领跑者的位置。2012年8月，思爱普推出了SAP BusinessObjects BI解决方案4.0版本的第三功能包，简称feature pack 3，随后又进行了改进整合。

以SAP HANA为基础，思爱普还打造了强大的实时数据平台，为用户提供全面的数据分析和处理服务。

☆甲骨文（Oracle）

自2009年收购Sun Microsystems公司（主要生产工作站和服务器）之后，甲骨文一直在进行硬件与软件的整合。该公司于2011年推出的大数据机(BDA)和Exalytics商务智能服务器，被认为是甲骨文强势进入大数据市场的标志。2012年初，正式供货的BDA和Exalytics预示甲骨文大数据平台解决方案的出台。

2012年12月13日，甲骨文宣布收购服务于石油、电气、供水行业的DataRaker公司，标志着大数据应用达到了一个新的趋势，开始向传统行业渗透，产生深入和全面的应用效果。

☆威睿（VMware）

威睿是全球桌面到数据中心虚拟化解决方案的领导厂商，它的虚拟化产品除了针对Hadoop进行优化外，还有围绕大数据分析和处理的项目。此外，Cetas和vFabric Data系列产品都降低了人们在进行数据处理分析时的复杂度。除了最为核心和拿手的虚拟化产品之外，威睿公司近几年也通过收购和自我研发推出了众多开源产品。比如HVE(Hadoop Virtualization Essential)的插件以及Serengeti的产品，都是威睿推出的开源的虚拟化产品。

☆Cloudera

Cloudera公司由来自脸书（Facebook）、谷歌和雅虎的前工程师杰夫·哈默巴切(Jeff Hammerbacher)、克里斯托弗·比塞格利亚(Christophe

Bisciglia)、埃姆·阿瓦达拉(Amr Awadallah)以及现任CEO、甲骨文前高管迈克·奥尔森(Mike Olson)在2008年创建。公司采用了NoSQL和Hadoop两种技术，由此获得了7600万美元的融资。

在2010年6月份，该公司正式推出了自己的企业产品。随后，Cloudera为其Apache Hadoop软件发行版增添了Cloudera管理器控制台及企业级的支持。现在它也与甲骨文进行密切合作，互相增加客户数量，推动彼此在大数据市场的份额。

☆MapR

MapR公司始终专注于可用性和数据安全的优化，它有自己的优势和独一无二的特性。比如，虽然和其他公司一样，MapR将基于开源的Hadoop产品商品化并进行销售，但它提供了很多不同于Hadoop的特性。它的产品为EMC的Greenplum HD企业版Hadoop提供技术支持。

不久前，MapR公司宣布了新的大数据平台MapR M7，这将为Hadoop与NoSQL提供更为方便、可靠和快速的服务。

☆Splunk

2003年成立并于2012年上市的美国商业智能软件提供商Splunk公司是公认的“大数据概念第一股”，它主要的业务就是向企业及客户提供数据引擎。它旗下的Machine Data软件的搜索功能具有强大的优势，而Splunk Free则专供个人用户使用，Splunk Enterprise则添加了支持多用户和分布式部署的功能。

在上述产品大获成功以后，Splunk公司随即又推出新的Splunk for Citrix XenDesktop解决方案，并在2012年的中旬将Splunk App for PCI Compliance 2.0全面推向市场。

谨慎：不是所有人都需要

诸如我们听到的、看到的和正在自觉或不自觉地参与的，大数据已成为一项大工程，它无处不在。我们对待它就像在迎接自己的终生伴侣，兴奋之情溢于言表。每个人都在想："嘿，大数据时代来了，我能从中得到什么好处呢？"从社交媒体、初创公司到北京的中关村，人们都在研究和部署大数据。

但是，正如前面我们提到的，大数据不是无源之水，你需要一个充足的理由来为它打开大门，让它进入你的世界；同时，你还需要为此付出不菲的代价。大多数公司缺乏预算，它们花不了大价钱来部署大数据技术解决方案，也请不起相关团队和大数据工程师。

大数据首先是一项产业，根据一份报告显示，2012年大数据带动了全球近300亿美元的IT支出，预计再过4年这个数字将超过2500亿美元。还有许多新兴国家难以预料的市场空间没有计算在内。要知道，这几乎是一个中等发达国家的全年国内经济总产值了。

那些使用大数据的辉煌案例到处都是，但距离某些特定人群总是如此遥远。比如，脸书的推广人员骄傲地说，他们每天要存储大约100TB的用户数据；美国国家安全局（NSA）每天要处理约24TB的数据。惊人的数字！确实令我们印象深刻。可是处理这些数据所需要的成本是多少呢？根据一项公开资料显示，NSA需要为45天的数据存储服务支付超过百万美元的费用，这个成本还在继续增加。在我几年的走访中，大多数公司的CIO也对我说，他们的预算支付不起大数据部署的成本。

所以，这是昂贵的门槛——公司如果想获得大数据服务，第一件要解决的事情就是提供充足的财务预算。

没钱？对不起，这不是卖白菜，也不是批发廉价商品或请几个经理人那么简单。因此我经常听到人们抱怨："大数据太贵了！"个人和企业都在仰天叹息，但同时又充满渴望。问题是，你真的需要它吗？

数据存储和处理的成本如此之高，成本变成了阻碍每一个人拥抱大数据的最大障碍，就像其他一切新生事物一样。以至于我们普通人——中小企业需要寻求其他的解决方案，让规模较小的公司和个体不被"大数据"拒之门外。

方案一：大数据的关键不是"大"。

大数据就一定"大"吗？虽然全球最大的科技公司都需要和PB级规模的数据打交道，它们当之无愧地成为对海量数据处理达到星级服务的用户。然而，我们的研究也表明，另外有95%的公司通常只需要使用0.5TB到40TB的数据，甚至更少。

脸书和NSA的故事并不能拿来作为普及版案例，它们不是常态。事实是，大公司的方案没有必要成为中小公司效仿的版本。在全美有5万多家公司的员工只有20到500人，它们大部分都有解决数据问题的需求，但它们并没有向脸书和NSA看齐，去建立一个成本高昂的数据帝国。

所以你看到，大数据市场最大的需求并不是那些居于世界前500强的大公司，而是排名在500到5万之间的公司。我们为何只关注那些极少数的例外，而忽视了普通的需求者呢？

将自己排除在PB级规模数据需求的用户之外，我们才有可能找到真正的方案。当大数据向我们走来时，我们应尽可能选择一个较小的接口，一样能享受同等的服务和便捷。

方案二：确定你是否真的需要它。

在向人们普及大数据时我经常在想，如果我们改变了大数据的定义，会发生什么？换一个角度，用更宏观的思维来思考它，你就能够跳出来，站在自我需求的角度去进行思考。

我们不妨这样考虑："大数据是一种主观状态，它描述的是一个公司（个人）的基础架构（现状）无法满足其对于数据处理的需求时的情形。"

从某种意义上来说，这个判断是"灰色"的，可能没有人们想象的那么灿烂美好。没有需求就不需要大数据。不过它更贴近事实：不是所有人都必须与大数据时代接轨，当你看到它扑面而来时，你要做的第一件事是确定自己是否真的需要它，然后再采取恰当的行动。

棱镜门背后的大数据革命

☆五个需要注意的问题

棱镜门事件爆发后，从斯诺登不断曝光的信息收集黑幕中，一夜之间人们突然意识到，数据的管理风险是无处不在的，不管是个人信息还是企业信息，在大数据时代都变成了极易被收集、窃取、分析和控制的“猎物”。这是因为对于大多数国家来说，大数据已经成为左右竞争局面的决定性力量，数据的安全风险也随之更加凸显。

● 强力机构可以轻而易举地利用技术收集我们的所有信息。

● 在今天，数据比任何时代更值钱，更能决定竞争的成败。

● 对每个国家、企业和个人来说，数据安全都已成为迫在眉睫的一项大挑战。

我们的国家和企业已经搜集并且存储了所有的数据，我们已升级了相关的技术，已为大数据产业的蓬勃发展奠定了基础。

接下来，中国人应该做些什么才能真正安全地发展大数据产业？我们应该如何对来自四面八方的个人、企业和国家数据进行保护？而且最为重要的是，我们如何才算是合法和健康地利用这些数据？

从企业和个人信息安全的角度来看，大数据有5个方面的问题需要我们注意。

网络安全

随着在线交易、在线对话、在线互动的兴起，在线数据越来越多，黑客们的犯罪动机也比以往任何时候都来得强烈。如今除了个人黑客之外，还出现了国家黑客，比如美国国安局。他们的组织性更强，更加专业，作案工具也是更加强大，作案手段更是层出不穷。相比于以往一次性数据泄露或者黑客攻击事件的小打小闹，现在数据一旦泄露，对整个企业、个人和国家而言，无异是重大打击，一着不慎就会满盘皆输，不仅会导致声誉受损、造成巨大的经济损失，严重的还要承担法律责任（比如金融机构的安全漏洞）。所以在大数据时代，网络的恢复能力以及防范策略可以说是至关重要。

云数据

云技术是新时代的技术产物，现在人们快速采用和实施诸如云服务时仍然存在大量的压力，这是因为我们对它们可能带来的风险和后果仍然没有办法预料和控制。尤为重要的是，云数据是黑客的目标，这是一个极具吸引力并能获取高价值信息的目标。因此，这就对企业制定与云计算相关的安全策略提出了极高的要求。

移动化

这个时代在变得“移动化”，人们对数据的需求增加，而数据的搜集、存储、访问、传输等工作都需要借助移动设备，所以大数据时代的来临也带动了移动设备的猛增。比如越来越多的员工用自己的移动设备进行办公，他们上班时拿着移动设备来到公司，下班后又拷贝了数据离开。我们不能否认，这很便利，有利于工作，也帮助企业节省了很大一笔开支，但也给企业带来了更大的安全隐患。要知道，移动设备是黑客入侵内网的绝佳跳板，比如以色列攻击伊朗核电站的手段就是靠一块很小的移动硬盘接入了核电站的工业计算机，从而释放病毒进行了致命攻击。移动化给企业的管理和安全保护带来了难度。

微妙而紧密的供应链

在今天这个全球化的时代，每个企业都是复杂的和互相依存的，都是全球供应链的一部分，但供应链本身恰恰是最薄弱的环节。信息将供应链紧密地联系在一起，从简单的数据到商业机密再到知识产权，而某一环节信息的泄露就可能导致整个供应链上的企业遭到巨大损失，甚至会违反法律，受到司法制裁。对全球化来说，信息安全是如此重要，它在整个供应链上扮演着血液的角色——试想我们的血液如果进入了病毒，会是什么后果？

隐私安全

随着产生、存储、分析的数据量越来越大，隐私问题在未来的几年也将愈加凸显。所以新的数据保护要求以及立法机构和监管部门的完善应当提上日程。

基于这些问题，大数据时代发生棱镜门事件一点也不意外，如果美国人不这么做才会让人觉得奇怪。事实上，不止美国人在这么干，不是吗？也不止各国的首脑在遭到威胁，不是吗？在这次震惊全球的情报丑闻中，有人看到的是“小丑出场”，但也有人已经看到棱镜门事件的背后是一场注定影响深远的革命，是一场与大数据有关的全球性变革。

在这场变革中，聚合以及大数据分析就像是营销情报的宝库，不管对国家、企业还是个人来说，我们都可能因此受害，更可能采取绝佳的方式从中受益。人们难以忽视大数据对我们的思维和技术革新的影响，比如营销和管理。

在这场变革到来时，我们应该保持敏捷性并在机遇出现前就做好准备，在机遇来临时果断抓住，而不是坐等时机错失再去亡羊补牢。当然，对很多国家和企业来说，一切都还处于初级阶段，而且目前也没有太多的外在要求来强制它提升数据处理技术以保证信息的安全。但是，企业每天处理的数据规模依然在保持增长，大数据分析并不会等你主动靠近它，而是会在全球化的作用下主动走向你。它使得商务决策越来越

依靠原生数据，使得信息的质量变得越来越重要。

同样，还有你无法忽视的数据安全。复杂的分析可以运用到相关的信息安全上，帮助你加固自己的数据仓库大门，以防成为强力机构随意提取的免费宝库。我们可以向全社会普及这些方案，鼓励中国企业和普通用户使用大数据分析来防骗和进行网络安全检测，使用大数据的社会分析技术以及多通道实时监测安全问题。

总的说来，大数据对中国具有无穷大的价值。我们从棱镜门事件中看到了它巨大的风险和安全隐患，但也体会到了一场山雨欲来风满楼的“鼎革之变”。这个世界是公平的，中国已错失大航海时代，但是上帝给了我们机会，让我们不会错失大数据时代。

☆全程展现——棱镜门的大数据背景

爱德华·斯诺登是前中情局雇员，他所披露出来的棱镜计划（PRISM），是一项由美国国家安全局（NSA）从小布什时期起开始实施的绝密电子监听工程。根据这项计划，美国的情报机构一直通过互联网公司进行数据挖掘工作，从音频、视频、图片、邮件、文档以及连接信息中分析个人的联系方式与行动。

监控的类型有10类：电邮信息、即时消息、视频、照片、存储数据、语音聊天、文件传输、视频会议、登录时间、社交网络资料的细节。其中还包括两个秘密的监视项目：

1.民众电话的通话记录；

2.民众的网络活动。

根据斯诺登的披露，从欧洲到拉美，从传统盟友到合作伙伴，从国家元首通话到日常会议记录，包括公民个人信息，美国情报部门都在进行监控。国安局可以接触到大量的个人聊天日志、存储的数据、语音通信、文件传输、个人社交网络数据。比如，美国政府要求电话公司提供数百万份的私人电话记录，详细到个人电话的时长、通话地点、通话双方的电话号码，把这些信息全部收集起来进行分析，然后选择重点监控对象。

有美国参议员证实，国安局的电话记录数据库至少已经有7年的历史。项目年度成本2000万美元。在2012年，作为总统每日简报的一部分，项目数据被引用高达1477次，可见它的使用频率。

美国政府辩称，阻止恐怖主义高于保护隐私权。奥巴马说：“你不能在拥有100%安全的情况下同时还拥有100%的隐私和100%的便利。”盟友英国也站出来支持说：“英国的守法公民永远不会知道政府部门为了阻止你的身份被盗或者挫败恐怖袭击所做的一切事情。”

听起来他们十分委屈，但用户数据实实在在地泄露了，而且里面

99%以上的人与恐怖主义没有半点关系。这些全球数据不但在用户自己的国家被存储，甚至也被传回了美国供国安局进行分析，比如脸书在它的隐私条款中称，所有用户必须同意他们的数据“被转送和存储在美国”。而《爱国者法案》给予了美国政府使用这些数据的权力。由于世界主要技术公司的总部都在美国，那么只要是接入了互联网或使用谷歌、思科等美国公司提供的联网设备，人们的隐私就可能被棱镜项目所侵犯，被转送和储存到美国政府相关的数据库。

斯诺登说，出于对隐私权的担心，他才采取了曝光行动。对此他坦言：“我不想生活在一个做那些事情的社会里，我不想生活在一言一行都被记录的世界里。”当他决然地迈出揭露行动的第一步时，我们才突然发觉，原来数据收集技术已经发达到了如此让人惊骇的程度。

☆现实——技术储备引发的变革

“棱镜”在技术层面是非常高端的大数据武器，是大数据革命的结果，这当然需要雄厚的技术储备才能实现。美国靠强大的数据通路和丰富的数据源收集大数据。已曝光的资料显示，美国有多个项目与棱镜数据的收集有关，比如从互联网骨干枢纽收集数据的“布拉尼计划”，从光纤网收集数据的“上游计划”等等。这些计划的成功都得益于一个现实——美国拥有全球电信最骨干的网络。

第一，带宽的技术优势。

由于美国网络基础设施建设走在了世界前面，导致北美跟欧洲、亚太、拉美之间的带宽远远地超过了其他的洲际带宽。相比从亚太地区直接发送数据到欧洲，经过美国中转会更快捷实惠，因此美国成为了数据传输的过路站，这给它监听数据提供了最基本的条件。

就像我们在现实中交换电子文档时，不一定靠U盘这种物理距离最近的方式，可能会采取MSN或腾讯QQ之类的即时通讯，因为它的速度更快、更方便，而且人们不会计较这些文件绕经网络服务器是否会被长期扫描和监测。这就是技术优势的巨大作用。

第二，数据传输诱导策略。

美国通过扩大数据传输带宽，可以诱导更多的数据流经本国，从它的家门口过，给它做一系列的数据截留分析工作大开方便之门。越多的数据流经美国，它能做的分析监测就越全面，那么它最终的收益就远远超过了风险。比如美国曾租用中国的卫星来传送数据，以提高非洲跟美国之间的带宽，就是一种付出较小风险获得较大收益的表现。

第三，控制数据通道，从其他数据源快速收集信息。

这一策略包括与电信运营商的合作和对其数据源的监控。例如，联邦政府的海外情报监听法庭要求美国电信运营商(Verizon)每天都要向国

安局提交元数据——电话记录数据，包括通话双方号码、通话长度等，虽然不包括通话内容，但已经提交了大部分的个人信息。

第四，通过与民营公司和其他国家建立技术联盟来收集和控制数据。

综上，美国最早意识到了需要加强信息管理与网络安全，而加强的办法就是用政府的手去控制民营企业及其他国家的相关服务商，建立技术联盟。

1978年美国国会通过了《外国情报监视法》（FISA），1986年又通过了《电子通信隐私法》（ECPA），1994年通过了《执法通信辅助法》（CALEA），从而建立起了全方位的保障体系，联邦政府可以从容地对本国及外国进行监听监视。

最重要的部分是与技术巨头有关的，即那些控制互联网的大型公司，在《执法通信辅助法》中规定，执法机关可以根据法院监听令直接接入电信网络，启动电信运营商交换机中的监听功能。这意味着美国法律要求电信运营商等网络、通信服务者必须为政府预留一定的接口以备不时之需。

根据《华盛顿邮报》的披露，在“棱镜”计划中，一共涉及至少9家美国的IT公司，微软是在2007年9月11日第一个加入其中的，苹果公司则是在2012年10月最后一个加入的。另外还包括思科、IBM、谷歌、高通、英特尔、雅虎、脸书和甲骨文等。它们几乎垄断了全球IT产业的所有领域，包含了从硬件到软件再到服务三个层面，自然为美国政府提供了强大的技术支持。

比如，如果你的联网电脑使用英特尔公司提供的某款芯片，就会发送一个序列号到英特尔公司，这也意味着在这台电脑上运行的一些信息也可以同时一并发送过去。另外，操作系统是网络软件运行的载体，联网后我们会经常收到自动更新的提示，这意味着垄断操作系统的微软公司可以轻而易举地掌握一台电脑的网络活动。同时，由于操作系统在不断更新，微软公司通常会最早发现其系统存在的漏洞，他们向政府安全部门提供的漏洞信息，会有助于情报机构攻击那些还没有修补漏洞的计算机。

再比如，人们用雅虎邮箱发邮件，用思科的网络电话通话，用谷歌的地图标注、搜索，用脸书发布社交状态，用MSN即时通讯聊天，所有这些网络活动，都会在各大公司的服务器上留下原始数据，而且还是人们主动提供的信息，自己花钱把信息送上门。

这些公司的服务器是如此之多，它们可以向美国政府开放直接访问的后门，帮助情报部门读取数据，甚至能够全程参与国安局的监控计

划。这些大数据的技术巨头，成为了政府收集信息和分析数据的绝好帮手，而这在民众毫不知情的情况下就可以完成。

海外盟友的数据来源包括澳大利亚、英国、日本、加拿大和新西兰等国家，比如著名的“五眼联盟”就是由美国、英国、加拿大、澳大利亚和新西兰五国组成，双方互通有无，协同收集数据。联盟成员甚至可以彼此监听对方国内的数据，绕开本方国内的法律禁区，然后交换数据。

第五，建立尽可能多的海外非盟友“数据源”。

当然，只有盟友数据源是不够的，美国还有大量的海外非盟友数据源。比如斯诺登就透露说，为了窃取中国大陆的数据，美国采取的办法是直接在中国境内建立数据源合作伙伴。香港中文大学在1995年成立了香港互联网交换中心，它的前身为港中大连接美国的数据专线，拥有服务于全香港的网络数据交换服务器。美国在这个基础上可以方便地潜入进来，对数据进行窃取。

为数众多的黑客也是这一数据源的提供者，国安局旗下有一个叫作TAO的机构，拥有多达六百名高级黑客，来自世界各地（包括中国）。思科公司提供的设备为这项工作提前留下了后门，尽管思科强烈否认这一质疑，但随着曝光的深入，否认的声音已越来越缺乏说服力。

只要能获得海量的源源不断的数据，美国强有力的大数据存储和分析系统就可以派上用场。

第六，建设大型数据中心来保存数据。

为了保存这些海量数据，还需要一个庞大的数据库和处理中心。NSA在犹他州耗资20亿美元建立了一个大型的数据中心，足以保存5000亿G的数据，相当于全球500年的通讯量。为了实现这一目标，NSA专门开发了一个叫作Accumulo的大数据存储系统，并与相关的有军方背景的民用公司合作，开发这一系统的商用版本，来持续获取数据利益。

第七，对元数据的挖掘技术，使美国有能力构筑关联图谱。

元数据是最基本的数据单位，在移动互联网快速发展的今天，我们每个人几乎每时每刻都在产生数据。比如姓名、电话号码、邮箱地址这些都可以称为元数据，它可以拿来当作节点，把有过联系的人、号码、邮箱用线连接起来，就构成了数据和信息背后的人物关联图谱。这表明，元数据虽然单个看起来不怎么重要，但大量集中起来，却非常便于构建个体之间的关联。

再比如，对电脑来说，元数据记录了一台计算机的工作环境，包括操作系统、浏览器、应用软件版本等基本信息，那么收集这些元数据，则是黑客发起网络攻击的必备步骤。对元数据的收集与分析能力，说明美国的网络监控水平已经具备了大数据时代的显著特征。有了这种对海

量元数据的存储与分析能力之后，这些庞杂的信息经过超级计算机的快速运算，就能从中显露出不易察觉的规律，从而为情报部门提供有效的情报信息。

在大数据时代，美国对于关联图谱的挖掘技术进展迅猛，使得从元数据中能够挖掘的隐私越来越多，简直到了无孔不入的地步。换句话说，现有的技术可以做到一切：侦测到你每天发送的短信数量、电话频率以及约会对象，并且深入地探查出你的全部生活和工作习惯，让你成为一个彻底透明的人。

一项调查显示，美国国安局拥有一个4.4万亿个节点、70万亿条相关联的图谱数据。按照全球70亿人口计算，国安局有能力为每人保存将近630种信息，可以分析出每个人多达1万种的关联。要知道，我们只需要4个时间点和位置就可以确定一个人的身份了，而且准确性已经高达95%，那么1万种呢？

这一技术的先进程度已超出普通人的想象，相关的技术告诉我们，即使你已隐姓埋名、流落天涯，只要他们有这个意愿，就可以轻松地找到你，并能穷尽你的一切社会关系，甚至比你自己知道的还要多。

第八，强大的分析工具：可视化和实时查询的大数据系统。

美国国安局还拥有一套大数据可视化和实时查询系统，名字叫作Boundless Informant。它的作用在于将监听、收集到的全球数据进行可视化和能够实时查询的分类，把不同的国家、地区用不同的颜色显示，构建出全球信息分布图。有了这样的大数据系统后，就具备了强大的分析能力，不管收集到多少数据，都可以轻松地整理、归类、分析和预测。

第九，拥有全数据挖掘技术，进行“无死角”数据收集。

美国一方面在挖掘元数据，另一方面也在发展对全数据的收集技术，争取做到数据收集无死角。像有一种叫作Narus的光纤监听设备，可以进行内容层监控，还有一种叫作爱因斯坦3号(Einstein 3)的系统，可以对数十亿邮件的全文内容进行扫描。语音监听和识别也是这一工作的组成部分，对多语种的语音视频内容进行分析辨别。

第十，为全球互联网的发展提供了可能性。

我们已经很难用一个特定的词汇来形容全球互联网的结构。但它的主要构成特点我们是清楚的，一方面它就像树一样，有根、主干和枝蔓，最后连接到每一个用户；另一方面它又是平状的，具有平等和无中心的特点，每一个节点的信息都可以在全球的网络中自由流通。也就是说从信息流通来看，互联网是一个扁平的世界，但在管理上，却仍然是自上而下的结构，有骨干网络来处理、管理所有的信息。

例如在中国，163和169骨干网承担着中国80%以上的网络数据流量，它们统称为中国公用计算机互联网（CHINANET），另外还有中国教育和科研计算机网（CERNET），全国科研机构的中国科技网（CSTNET），中国金桥信息网（CHINAGBN）。它们共同组成了由上到下的中国四大骨干网络。从网络监控和攻击的角度来说，自然是从上往下更好，这有利于获取更多的信息，也拥有更大的控制权限。

从棱镜门事件我们可以看到，由于具备足够的资金、技术和不受限制的权力，较大的机构是大数据的最大受益者，它们可以窥探个体的信息和收集、预测人们的关联数据，达到控制人们需求并实现一系列组织计划的目的。要想充分防范大数据技术的滥用，就需要发挥我们每一个人的创造性，由人去主导大数据的进程，而不是成为数据可以控制的一部分。

我们每个普通人都需要参与进来，思考大数据技术的合理开发，并加入这个神奇的新时代，成为它的主人。

CHAPTER 2 大数据时代给世界的巨大转型机会我们能做什么，我们要做什么

● 不管怎么样，我们现在唯一要做的事情是，张开双臂，积极地迎接大数据时代。

● 今天，如果你正在或者打算尝试使用云技术，比如云分享或云计算，那么恭喜，你已成为大数据时代的一员，或者是它的受害者。

● 大数据的出现首先是一种机遇，其次它带来了重大的挑战。我们既要享受它产生的福利，也要警惕它背后潜在的弊病。

☆先抓住它的核心问题

大数据具有多层结构，这意味着它的形式多变，类型也很丰富。有人认为，人们越来越频繁地使用互联网进行搜索是形成数据多样性的主要原因，这当然是有道理的，但最主要的还是由于新型多结构数据的出现，以及包括网络日志、社交媒体、手机通话记录及传感器网络等数据类型形成的。数据传感器可以安在更多的地方，比如汽车、飞机、卫星或手机上，都增加了数据的多样性。

和传统的业务数据比起来，大数据又存在不规则和模糊不清的特性，因此人们很难甚至没有办法使用传统软件或方法进行分析，有时就连收集也成为一种不可能。随着传统的业务数据的演变，它的格式已能够被标准的智能软件识别，目前我们面临的挑战是处理并且从以各种形式呈现的复杂数据中挖掘价值。

一项关于数据创建速度的调查显示，到2020年时，全世界将拥有220亿部互联网连接设备。在大数据时代，数据被创建和移动的速度是非常快的，创建实时数据流是一种流行趋势，因为有高速电脑和服务器的存在，这不是什么难事。在这个基础上，我们还必须懂得如何快速处理、分析数据并满足用户的实时需求。

我们（包括企业和个人）面临着数据量的大规模增长，这是一个不争的事实。再过15年，全世界的数据量将扩大到今天的50到60倍。它的规模是一个时刻在变化的指标，谁也无法预计将来还会出现多大程度的技术飞跃。但可以肯定的是，数据量的增长只会越来越快，绝不会放慢。另外，各种意想不到的来源都能产生数据，也都能保存数据。

☆想想你能做什么？

在将来，我们的竞争优势（超越强手的优势）将来自何处呢？想想这个问题，你就明白了大数据赋予人类的使命。未来的竞争优势已很难从制造业或工业资源的“仓库”中提取，而是来自于数据，还有相应的收集、分析和使用它的能力。

在未来的大数据时代，只有能够提供功能最为丰富、数据量最大的数据平台的公司才可以在企业的竞争中获胜；只有能够拥有最强大的大数据产业的国家才可以在国家的竞争中笑到最后。

大数据科学家舍恩伯格说：“现在有越来越多的数据，人们可以收集、分析与所要研究的问题相关的更多信息。通过这些数据，人们能够得到很多的洞识，帮助他们做出选择与决策。”

他认为，只有我们分析了所有的相关现象、所有的数据或大多数的数据，才能够发现以前没看到过的问题与选择。因此，人们必须学会善用更多的数据。在这个大的前提下，舍恩伯格为我们指出：大数据时代最大的转变就是不再强烈地渴求因果关系，而是更多地去关注相关关系。（我对这一观点持有异议，后面的篇幅我们会有重点讨论。）

也就是说，舍恩伯格认为，在大数据时代，我们只要知道了“是什么”，不需要知道“为什么”，就能达到更宏伟的目标。这是全新的思维，也正是我们要做的事情。我们必须创造新的交流方式，必须建立新的认知，才能跟上大数据的步伐，成为新型的现代人。

☆认清数据的价值：重复使用

数据的价值是什么？关键的一点是，它总在改变，从不是固定不变的。在以前（小数据时代），数据往往使用一次就失去了意义，但在今天，数据却可以重复使用。你可以随时调取它、使用它，不需要担心它损坏或失去功能。

真正价值就在于它可以一而再、再而三地使用。这种“再使用”的价值让数据的重要性比过去陡增了几百倍甚至成千上万倍。

由于这一新的特点，互联网的作用被无限扩大了，并最终催生了遍布每个行业的大数据产业，因为人人都有重复使用数据的需求。企业有，个人也有。对整个世界来说，这可能意味着大数据产业将引领经济的发展，全方位地影响我们的生活。

☆到了为自己建立大数据时代思维的时候了吗？

但是，对普通人而言，我们需要做些什么才可以更适应时代，或者才能够走在这个时代的前沿？

仔细想一想，你有机会来引领属于自己的大数据时代吗？

在美国有一家创新企业德克德公司，它可以帮助人们做购买决策，告诉消费者什么时候买什么产品，什么时候买最便宜。它总是能够精明地预测产品的价格趋势。

它是如何做到的？背后强大的驱动力就是大数据的支持。他们在全球的网站上搜集到了数以十亿计的数据，然后帮助数以十万计的用户省钱，为他们的采购找到最好的时间，提高生产率，降低交易的成本，为

那些终端的消费者带去更多价值。

在这类模式下，尽管一些零售商的利润会进一步受挤压，但从商业本质上来讲，可以把钱更多地放回到消费者的口袋里，让人们的购物变得更加富有理性，不至于花大钱办小事，并可以降低自己购买假货的概率。

这是依靠大数据催生出的一项全新产业。这家为数以十万计的客户省钱的公司，在不久前，被一家超级企业以高价收购。

另一个例子与SWIFT公司有关，它是全球最大的支付平台，在该平台上的每一笔交易都可以进行大数据的分析。他们可以预测一个经济体的健康性和增长性。比如提供世界某一个地区的经济指数，你可以实时实地得到对不同地区的精确统计、计算与预测。

数据可以告诉我们每一名客户的消费倾向：他们想要什么？喜欢什么？每个人的需求有哪些区别？哪一些可以整合到一起进行分类分析？具有超前眼光的公司早就据此布局，实现了对消费者和用户的数据化分析、服务与预测。

多数人没有能力去创办一家这样的公司，但我们可以在大数据产业中发现自己能够饰演的角色，例如数据工程师、提供思维或开发程序的人，当然还有收集与整理数据的人。我们在生活中就可以顺利地建立这种思维，成为一名当之无愧的“数据控”，打理好自己的生活。

☆我们的未来——开发与充分利用数据

你可以仔细想一想，数据的收集、分析和处理，应该是怎样进行的。我们将按照顺序来逐一介绍和讨论，并提出与一些广为流行的常识有所区别的观点。

第一步：数据的收集。

收集是大数据供应链的第一个环节。数据是大数据产业的原料，没有原料，任何产业都没有办法发展。从广义的角度，信息就是数据，我们可以通过各种公共或者私人的渠道获得信息。这些信息各式各样，来自不同的地方，都被我们汇集起来。

随着收集数据的成本越来越低（因为汇集数据的市场日益发达），我们用比较低的、能够接受的价格来获得几乎所有的宝贵数据都是有可能的。这些信息包括一切领域，甚至是你穷极一生都无法了解的人类文明史的全部学科——从社交网络、情感、军事政治到天气预报、经济指标和乏味的公共信息，如今都成为了我们的“大数据加工厂”的原料。

你可以从互联网收集信息，点击鼠标到达任意网站，查看你感兴趣的东西，然后记录；

你可以从智能手机、iPad或其他移动数据平台收集信息，它们总能

根据你的喜好忠实地为你提供信息服务；

你可以通过邮件或流量统计工具收集信息，这是与特定组织相关的数据。比如消费者的访问量、产品召回度和顾客的忠诚指数等，你都能付出极低的成本获得它们。

既然收集成为了轻易可达成的目标，那么在技术条件允许的情况下，合法性的讨论就被提上了桌面。“我可以随便把信息拿过来吗？没有限制吗？”当然有。某些数据会受到严格的管制，比如医疗信息、个人房产和婚姻信息。在不同的情况下，收集信息可能面临合法与非法两种判定。如果你涉及对个人身份识别信息的利用，就可能非法；如果不是，则存在法律的模糊区域。

在世界范围内，我们的司法系统对于网络信息是否代表个人身份（隐私）的判定并没有统一的意见，这包括IP地址。但是，最近美国的一些地区法院已开始立法约束，明确了一些管理条例，比如加州的最高法院裁定邮政编码为个人信息，对相关数据能被哪些机构收集做出了强制性的约束。

在电脑和网络普及时代，每个人都成为潜在的数据来源。就拿手机来说，进入智能机时代后，手机成为绝佳的信息采集和发送装置，它可以感知光线、声音、动作、位置，附近的网络、电脑、其他手机（使用人及其位置）等。这是理想的数据采集器，手机使用者如果安装了厂商的软件，就自动加入了数据供应链。有时他们对此缺乏认知，因为人们更多关注的是使用功能和便捷服务（包括软件升级和信息获得功能）。

这意味着抛开合法与非法的判定，信息正变得海量和无处不在。要达到匹配的收集速度，是一项极具挑战性的工作。要完成这个工作，我们就需要使用新的技术和平台，促进技术革新，从而推动一系列产业。

第二步：数据的提取和清除。

数据收集好，不意味着就万事大吉了。恰恰相反，工作才刚刚开始。收集好了，就必须把它们提取出来进行整理分类。在情报领域，这被称为“提取、转换和加载”，要把数据存进一个设计好的数据库，进行一定处理，然后才易于调取和使用。

大数据的一个最显著特征就是非结构化。它不具有天然的结构性，信息在收集好的最初阶段往往是混乱的、杂乱的和缺乏规律的，什么来源和性质的信息都有。这表明我们在提取和分析工作开展前，并不清楚这些信息的内在架构。

很头疼是吗？接下来，对信息转换的需求出现了。我们需要在保持源数据的同时，又能快速地分析数据，把不同的结构定义出来。

第三步：硬件的发展。

这时，硬件的发展就被提上了日程。没有升级的硬件，就无法承载升级的软件，也就不能满足庞大的分析工程。我们收集、提取的任何数据需要经过人或机器的分析，更多的还要靠机器而非人。

在这里，硬件是以计算、存储和联网的形式存在的，多以电脑为载体，成为数据服务器的一部分。大数据并不会改变这一点，但是它改变了传统硬件的用途，也使云计算成为了宠儿。因为云计算使得数据虚拟化和实时化，既可以接纳海量数据进行分析，又能随时清除这些数据，做到按需分析，这使对海量数据进行精确分析成为了可能。

第四步：平台的重要性。

我们要创造可用来快速处理海量信息的平台和框架，没有这个平台，前述工作将变得不可能。在这个平台上，我们加快数据分析的方法就是将数据分解，再对若干部分分别进行分析。当然还有另外的途径，即建立一个文档处理步骤的路径，每一个步骤都对特定的任务进行最优化的分析。

平台还要具备一个重要特点：迅速出结果，而不是只能处理大量的数据却无法保证实时性。这一点相当重要，因为人们既需要实时信息，又需要反复地分析这些数据。比如提供网络搜索结果，百度不可能在24小时后才显示搜索页面，必须瞬间呈现才能满足用户需求；航班、酒店信息等也必须实时呈现。实现这些目标的唯一方法，就是平台具备分派任务的功能，这就是为什么大型网络公司都有上百个服务器。最后，平台也要满足人们反复使用的需求，这对技术的要求更高。

第五步：机器智能。

在大数据供应链中，机器的智能相当关键。因为数据太多了，无法用手工处理。特别是对于今天我们想要分析的大部分数据——整个大数据产业来说，离开了机器的帮助寸步难行。机器的智能化是必然的趋势，谁占领了机器智能的最高阵地，谁就在大数据产业中占得了先机，拥有了核心技术，就不会受制于人，而会达到“制人”的境界。

在数据和信息的收集、提取阶段，机器就已经介入帮忙了。比如，对大量信息进行推导，归纳出数据的含义；对几千名客服每日、每周的服务满意度进行总结；对车票、机票的订票量进行统计。你不可能让人工参与其中，因为他们太慢，满足不了实时性的要求。

机器不但参与其中，它的学习能力也很重要。如果我们要分析信息，就要试着在更高难度的环境中尝试更快的速度，自然就要不断提升机器的智能。换言之，在大数据时代，我们的机器将越来越聪明。它们会逐渐变得可以更深入地思考，拥有一定的情感模式和逻辑判断力。虽然我们还无法预测智能机器的未来，但它们已经表现得像人类智慧的初

期阶段了。

第六步：人类的作用。

虽然机器的智能对数据分析相当重要，但是永远不可能替代人类。人的眼睛、耳朵和大脑仍然（可能是永远）是这个世界上最智能的工具。机器不管如何进步，最终都只是为了延伸视觉的维度，以人类可读的形式提供数据。

所以，重要的不是机器或人一方，而是“人机互动”。大部分的分析师都清醒地知道，人是数据的主宰者，机器只是一个打工仔。凯瑞尔（Creve）是人机互动研究的先行者，他设计出了利用几十个独立数据源的系统，功能十分强大，不但能在可操作的3D环境下对系统进行显示，而且能辅之以声音和其他信号。他的研究表明，如果人们用这种方式输入数据，分析员不用花几个小时，而是只需几分钟便能寻找到答案。

人类的作用在于控制机器，成为数据的主人，在此基础上提升人机互动的速度和并行性。当然，人类还需要给机器设计新的界面和多重感应环境，以方便数据分析师和机器一起埋头苦干，高效地处理数据。

第七步：数据的存储。

我们必须考虑数据的存储。实际上这个问题在一开始就会成为人们重点设计的环节，因为大数据所占的储存空间实在太大了。

在庞大的数据中，除了一些源信息，还存在着大量的已作了改变的数据。我们收集、整理、改动、加工它们；另外也有通过分析得出的简表和表格，并由此产生了许多格式文件。为了尽可能多地提供空间，我们要研发新科技，让数据拥有更宽敞的“家”。

通常来说，储存是指什么？一位数据专家说：“储存就是使用传统的平面文件和相关的数据集加上后结构化查询语言（post-SQL）储存系统对云数据和初始数据进行保存。”如果在大数据供应链中缺乏这一环节，我们就无法备份所有东西，数据库就难以达标，不能支持庞大的工作量。这就像一个人虽然饥饿却只有很小的胃一样。

第八步：达成分享数据和协同行动的目标。

这关系到数据分析的目的——有效的行动。数据要体现分享和行动的价值，即便不能全世界共享，也要在有限的圈子里做到资源最大化，提升数据资产。就像好的企业利用大数据做出运营决策，无论是雇佣或辞退，或者战略规划、市场信息定位等，都是数据分享和行动的绝佳体现。

只是研究数据而不进行根本改变，不但得不到有力的结果，反而会使自己成为大数据变革的阻力。在每次发生重大的变革时，必然会出现

一些似曾相识的拒绝变革的行为，所有的行业、领域都是如此。科技史上，大型主机、客户服务端计算、互联网等新技术的诞生，无一例外都受到过诋毁者的攻击，几百年的工业革命也遭遇过贵族的质疑和拼死阻挠。大数据和云计算也未能幸免，它们时刻面临类似的人为因素。

他们最喜欢说的一句话是：“未经明确验证这是可行的。”

但事实上，反对者缺乏大数据思维。大数据思维在本质上是一种实验性思维，需要的是分享精神和行动力，需要快速行动而不是犹豫不决，需要果断而非优柔寡断。大数据是一场新的席卷全球的大运动，提倡的是快速、反复地学习并与用户紧密联系，惠及每一个使用者。

第九步：测量并且收集汇总反馈的数据。

这是最后一步，但也是新的开始。大数据从反馈开始，也总是与反馈有关，获得信息，发送信息，再获得信息。

如果你只能分析数据，这没什么了不起的，并没有什么过人之处。为了使整个数据工程（长久有益的部分）流畅持久地运转起来，你必须从数据分析结果中选择一条行动路线，之后观察究竟发生了什么，就像商品的售后部门所做的工作，利用这一信息收集新数据或者采用不同的方法进行分析，看看如何改进之前的工作。

这一步代表着一个持续优化的过程，它将影响到大数据使用者的方方面面。你不能只是将收集反馈的工作当成一顶漂亮的帽子戴在头上。成为摆设是危险的，帽子不但要好看，还必须暖和，然后去设计更好的帽子，改进自己的“获利”能力，这是大数据的终极目标。

数据正在影响着整个世界，这已成为一个事实。之前几年我们还发现有很多的不向数据分析开放的市场，比如文化出版、音乐、房地产等行业，但现在那道森严的壁垒也被打破了。我们现在发现一切行业都在向数字化、数据化前进，中国内地的中小企业也在融入全球供应链，成为大数据时代的一部分。

但是，我们要以数据为世界的灵魂吗？当然不是。在我们走向大数据时代的每一分钟，都要思考这个问题，这有利于保持我们“人性”的清醒，而不是将世界完全托付给数据，哪怕它是超级智能的。

反馈经济——新的营销模式在兴起

最近几年来，反馈经济（Feedback Economy）成为经济学界研究的重点。事实上，在很多领域，我们已经看到了反馈经济的兴起，它代表着新的营销模式和全球产业转型的推动力。而这正是大数据送给世界和中国的第一份经济礼物。

一切以数据为中心、在数据的基础上进行分析和持续的优化，不仅简单地提高了企业经营效率，而且还能让我们做好准备迎接更大型、更重要的改变。手机和电脑等设备记录着我们的数据，比如地点、喜好、习惯、状态等，把它们时时地记下来，通过移动互联网，将上述大量的数据传送到云中心，然后被商家收集和分析。

商家利用云计算技术来处理这些数据，进行比较，做出决策，最后反馈到个人移动设备终端或其他设备上。这样，数据流就形成了一个“闭环”，不断地更新和反馈，提升产品质量，提高用户满意度。固定不变和循环往复的反馈，将导致效率提升与最优化的积极结果，这成为了商业公司和政府部门的运营标准，也成为了反馈经济的基本特点。它将最终超越信息经济，信息本身没有价值，只有进入这个闭环之后才变得富有价值。

2013年，有三名斯坦福大学的学生创立了一家公司，专门防治脊椎病。人们平时的坐姿都有问题，这是导致脊椎病的主要原因，但对坐姿问题，多数人一般意识不到，即便知道也很难保持长时间的注意力。

怎么解决这个问题呢？

这家公司提出的方案是，设计一个传感器，把它放到腰带上。传感器可以时时地监测我们的坐姿，并且通过移动网络，传输到一个云中心。云中心的服务器不断地积累数据，通过比较分析，评估你的坐姿并且计算出多长时间就需要调整，然后它将需要调整的信号发送到你的手机上。就像闹钟一样，时间一到，它就提醒你。

更强大的功能是，它还可以把你的坐姿状态的数据发给你的好友和亲人，一起分享和互相提醒。比如，一名热爱学习的年轻人戴上了这个传感器，不但他会收到提醒，他爸爸妈妈也可以收到这个提醒信息，然后过来敲他的房间：“喂，你该换个姿势了。”假如你的很多朋友都带上了这个设备，那么这个云中心的数据就更加丰富了，人们在这个中心平台可以看到彼此的状态，纠正对方的错误或者学习对方的经验。

这个故事只是“反馈经济”形态的一个很小的展现，现实中有更多形式的反馈经济的企业。这只是一个良好的开始，如果人们愿意，我们每个人在自己的一生中产生的全部数据都可以记录下来，为现实应用提供

支持。

人类的个体行为、群体行为、社会行为的调整与改变都是非常困难的，这不但与习惯有关，还和缺乏足够精确的数据记录有关。中国人更是如此，从古到今，我们一直缺乏足够的精确的数字意识，比如人们问对方什么时候可以完成工作时，得到的回答通常是“差不多了”“快了”而不是“5号”或“还有两个小时”。建立精确习惯比较困难，但是，利用反馈经济中的移动互联网、大数据和云计算技术，我们就能让自己的行为变得可记录、可计算、可模拟和可反馈，让每个人都参与到大数据的积极价值中来，促进经济的转型。

虽然移动互联网在时时传回人体行为的数据时，它的数据总量是庞大的（据统计，人的一生行为产生的数据大小约为1000T（1T=1024G）），但是云计算技术和大数据的存储能力可以解决这个问题。从技术的角度，人的行为从此变得可记录和可分析。大数据实现这一点非常轻松，几乎毫不费力，这已奠定完美的技术基础。在分析完人的行为的数据之后，云中心可以反馈给每个人，然后根据个人的喜好互相分享。同时，企业也可以利用这种模式，将自己的产品、服务和一切信息发送给用户，从他们那里得到反馈数据，并提升自己的经营效率。

反馈越及时，结果就越有意义。个人的反馈和分享会带动群体的反馈和分享，从而使得我们的社会和经济开始普遍受益。比如，随着数据收集能力的加强，云中心计算出来的模型（分析和预测结果）越来越接近于人的行为本身的状态，也就越来越精确，意义也就越大。当它在各个领域获得推广后，就会产生成倍放大的效应。

教育——可以根据学生的成绩、兴趣和行为，找出他们的天赋进行因材施教；

售后——可以根据用户使用数据收集、分析，获得反馈，设计更好的产品或服务；

医疗——可以记录每个人的身体压力、机能、精神状态，做好预防和术后恢复工作；

…………

这样一来，反馈经济的规模将会越来越大，实现产业转型，成为新时代经济的主角。从这个角度说，谁在未来掌握了足够的用户数据，谁就赢得了明天。今天，中国人需要思考的是如何借助这一正在席卷全球的大趋势，搭上开往反馈经济的列车，而且争取领先，成为这一新时代的引领者。

为了搭上这趟列车，传统的行业应该怎么办？唯一需要做的，就是

结合自身的情况，在合适的时候采用大数据技术，加入到反馈经济的阵营，用反馈思维来带动企业的进步和腾飞。从现在开始，人们要做的就是为迎接反馈经济进行必要的准备工作，打好基础，在时机到来时无缝接入。

我们要赢得一场反馈经济的战争，需要具备以下三点：

首先，能更好地收集并且分析信息；

其次，能更快处理这些信息并将所掌握的信息融入到下一个反馈的循环中；

最后，建立一个运行有序和有效的“闭环”。

美国一位研究中国经济的学者约翰·克伦对我说：“中国人或许首先要认识到，进行数据减肥比吞食更多的数据来得重要，因为中国已经形成了一个信息肥胖的社会，并不是收集全部数据和分析全部信息就能得到很好的反馈，取得计划中的效果。”

克伦的深刻见解是如此重要。在过去的30年中，中国人的人际互动已从现实世界转移至了网络世界，信息化来得是如此之快，以至于每个人都拥有一部智能手机，每家公司都能轻而易举地获得海量数据，虽然这和“精确分析”画不上等号。人与人之间的相互作用变成了数字化形式，互动总是同时发生，且极易复制。

比如中国人习惯了在淘宝购物，先看一遍用户评价，再敲几个键，两三天的时间就能收到货物；如果想把自己的心情通告给全世界，在微博上写几十个字就可以了，比打电话找几个朋友倾诉更容易；因为电子邮件沟通太慢，中国人交流问题几乎从不使用它，而是选择即时通讯软件。于是，世界被数字化了，我们被淹没在了数据的海洋中，同时也构筑了一个新的营销市场。这是反馈经济兴起的基础。

从现实世界到网络世界的转变，将所有行业的摩擦系数减少到了零。未来，这一进程仍将加快，直到所有人都乐此不疲地享受“反馈”带来的益处。我们需要采取更好的办法来观察和适应新的环境，来获得自己的机会。在转型面前，作出正确的决定、付诸行动和进行实验，把已经学到的知识应用至未来的行动中，成为大数据时代经济转型的赢家。

数据大爆炸——信息过量

互联网技术飞速发展的直接结果，就是我们生活的这个世界每天都会出现大量的信息，而且它的增长速度是一件近乎恐怖的事情，因此我们称之为“信息爆炸”（Information Explosion），又称为信息过量。它的冲击波像爆炸一样席卷整个地球，没有人可以抗拒。

对这一现象最早的总结出现于20世纪80年代。统计表明，在1980年到1990年之间的十年，全球信息量相比过去以惊人的速度飞增，平均每20个月就增加近一倍。人们感到信息的扩增就像炸弹的气浪，令人室息。进入90年代后，这一现象更加明显，随着互联网在90年代末的出现，信息极度膨胀，人们开始明显感觉到自己已经进入了一个“数据过剩”的时代。

互联网一个很大的特点就是它使信息的采集、传播的速度和规模达到了空前的水平，并且它的互联性实现了全球的信息共享与交互。与此同时，现代通信和传播技术也极大地提高了信息传播的速度和广度，比如由卫星通信、计算机通信和电视传媒等技术手段形成的微波及光纤通信网络，克服了传统的时间和空间障碍，把世界更进一步地连为了一体。随之而来的就是信息汹涌而来，多到让人无所适从。

从浩如烟海的信息中迅速而准确地找到那些自己最需要的信息，就成了一件比较困难的事情。这时，人们在对信息极度渴求的同时，也在怀疑信息。因为我们即使每天24小时看这些信息，恐怕也阅读不完。更何况，其中存在着大量的无用和不真实的信息，信息源也良萎不齐。它们增长的速度远比我们理解的速度要快，并且像海浪似的从四面八方涌入我们的生活。

信息过量主要表现在五个方面：

1.新闻资讯飞速增加，且逐渐具有实时性，即在地球另一端发生的事件，我们可以第一时间看到直播报道。

2.娱乐信息铺天盖地，并且真假难辨。

3.广告信息充斥生活，遍布每一个角落，而且更新迅速。

4.科技信息飞速递增，科技进步的速度超出了人们理解的速度。

5.个人的接受能力严重“超载”，人们每天都在为信息过量感到痛苦，这导致了一系列问题，同时也引发了大数据革命，因为我们需要一个有效的工具来帮助处理信息。

☆过量信息的来源

1.科技的传播速度增加，使以前没有发现或无法进入人类视野的信息，在今天也能够迅速地展现在我们面前。比如，现在我们每年出版的

图书达到几十万种，推动知识老化的同时，也在加速知识更新和更广的传播力度。在新的传播技术的推动下，全球印刷信息的生产量每5年就会翻一番。最近30年，人类生产的信息已经超过了过去5000年的总和。

2.信息管理不善导致的“传播失控”，产生了大量的无用与虚假信息，造成了信息环境的污染和“信息垃圾”的产生。比如现在的网络平台，任何人都可以自由发表意见，随时发表看法，且付出的成本几乎是可以忽略不计的。那么每个人都成为了一个全球范围内的信息发布者和数据的制造者。一方面这方便了人们从不同渠道获得信息，另一方面也增加了人们利用信息的困难，因为真假难辨。

3.病毒会产生错误数据或误导性的信息，也会造成大量的垃圾文件并消耗人们的精力。

4.同一件事在不同的角度、不同的视角得出了不同的描述与反映，就产生了大量的雷同但有区别的信息。这些信息虽然指向同一事物，但内容可能大相径庭，这让信息的使用者难以抓住“真相”，沉溺在这些不同的信息源中自相矛盾。

5.不健康信息，比如色情网站和黑客站点，以及存在不良目的的信息发布平台。

☆数据大爆炸——已经多到不能计量

信息多到了什么程度已经没有人说得准，因为今天的预测和统计在明天就会以更快的速度更新。我们每天从工作到生活，无时无刻不在制造新的数据，各个行业与机构每天也在不停地收集、传递、储存庞大的数据，全世界的数据量每两年就会增加一倍，未来可能几个月甚至几天就会增加一倍（这是必然会实现的，而且将很快实现）。

过量的无法计量的数据，已经多到让人们抵消了对数据质量的要求，会觉得拥有这么多信息总能在里面找到自己需要的东西，但事实可能并非如此。

牛津大学的维克多教授说：“资料仅仅是真相的幻影，更多的数据并不能引导我们发掘更多的真相，相反它只会引导出更多的数据以及更多的问题。虽然过量的数据可以增加我们的洞察能力，让我从中找到‘是什么’，却不一定能够找出‘为什么’。”

维克多认为大数据是自印刷术以来人类社会最大的革命，过量数据引发的一系列变革，已经改变了我们的工作、生活与思维。随着相关技术的发展，原来仅限于情报机关和大型企业的数据关联与分析技术，将会越来越普及，应用在商业、行政、科学、医疗各个领域，使得分析后的数据成为最宝贵的资产。

数据的数量不等于数据的质量，所以数据在收集之后必须进行整

理、分析。因为由于数据来源的零散、没有结构、没有规划、没有固定目的，导致即便数据再多，如果盲目用在特定的目标上，也必然产生缺乏质量的问题。

只有经过严密的富有逻辑的整理、分析、关联，才可以作为预测的根据。就像当当网分析了你的喜好后才能给你推荐你可能喜欢的书，百度分析了你的搜索习惯后才能在搜寻的结果页面展示你可能感兴趣的广告。无论是公安部门对犯罪多发地区的巡逻布置，保险公司对风险的预测，还是气象部门对未来15天的天气情况的判断，都是大量数据分析的结果。

☆不利和有利影响

信息过量的同时也会导致信息匮乏。因为从数量的角度来看，信息过量是指由于传播技术的进步以及传播环境的日渐放开，信息呈现海量级的涌现，为大数据技术提供了“原材料”支持。但由于信息太多而受众的分辨能力有限，无法获得最需要的信息，不能满足对信息的真实需求，就又产生了信息匮乏。这种匮乏是相对性的，受众面对鱼龙混杂的海量信息，不知所措。真正有价值的数据被大量的垃圾信息淹没了。

实际上的情况则是，我们一方面享受着丰富信息带来的便利，另一方面却在同时忍受着数据爆炸的困扰，毕竟并不是每个人都有大数据技术的需求和对海量数据的需要。但为了应对信息过量的负面影响，我们仍然不得不提升自己处理信息和分析信息的能力，以提高自己的决策效率。从长远来看，即便在当前付出一些高昂的成本，它也是非常有必要的。

全球产业链正面临大调整

在一次管理课堂上，有人让我解释到底什么才是产业链，尤其是全球产业链。因为人们每天都在听到、看到这个词，但很少有人能用通俗易懂的文字把它的内涵讲明白。

☆产业链的基本概念

产业链是经济学中的一个概念，简单地说，有两个关键词，一个是分工，一个是授权。就是各个产业部门（同一产业中的不同公司）之间基于一定的技术经济关联，并且依据特定的逻辑关系和时空布局关系而客观形成的链条式的关联形态。它是一个包含价值链、企业链、供需链和空间链四个维度的概念，从而形成了一个稳定的对接机制，就像一只无形之手，来调控产业链条中的每一个环节的角色。

在产业链中，存在着大量的上下游关系和相互价值的交换，上游环节向下游环节输送产品或服务，而下游环节向上游环节反馈信息。例如，苹果公司设计产品和提供服务（售后），代工厂则得到授权，替它生产产品并反馈相关的信息，在这个链条上，就构成了从高端到低端再到用户的一个基本配置。

在本质上，这就是一个具有某种内在联系的企业群结构了，既有了结构属性，也有了价值属性。授权方永远在上，被授权方永远在下。

☆全球产业链的分工本质

全球产业链的本质就是国际分工。它当然不是由抽签决定的，而是依据国家实力的高低，根据技术、竞争能力的不同，来区分每个国家的角色。

现在的国际分工具备两个主要特点：

1.由强国和其所属大型的跨国公司来主导国际分工，弱国和中小企业处于被动和受支配的地位。这当然会导致世界各国的发展利益分配失衡。比如，中国的许多外贸企业只能从事加工和组装生产，付出最辛苦的劳动，但只能获得不到10%的总利润，其他利润被跨国公司拿走。就像一件耐克品牌的衣服，中国获得授权的服装企业进行贴牌生产，一件衣服总利润100元的话，中国企业只能得到10元，90元被授权方耐克总公司拿走（它什么都没干，只是授权而已）。

2.生产和消费分离，贸易的数量与实际所获的利益也分离，从而导致相关国家（主要是产业链中下游的国家）出现国际收支不平衡的现象，进一步掠夺利润，加大差距，巩固强国和跨国公司在产业链中的高端、上游地位。

这种分工的本质，就是强者为王，谁拥有技术和知识产权，谁就掌

握了规则。在生产全球化进程中，随着全球产业结构的大规模深刻调整，发达国家不断地加快自身的产业升级并且优化增长的方式，然后本国的产业结构向知识密集、技术密集和服务密集的方向升级，增强了自己的产业竞争和技术优势。与此同时，把生产能力转移到发展中国家，两者变成一种全球共享型的生产模式，类似于老板和打工仔之间的合作。

到今天为止，世界范围内已经完成三次产业转移。分别是20世纪五六十年代美国把淘汰的钢铁、纺织等劳动密集型产业转移到日韩等国；20世纪七八十年代日、韩、美等国将劳动密集型产业转移至东南亚发展中国家；20世纪八九十年代随着中国的改革开放，欧洲及美日等发达国家又将劳动密集型企业转移到了中国的东南沿海，就是我们常说的长江和珠江三角洲地区。

☆产业转移：中国的机会

现在，产业转移仍在进行。最近十年来，随着经济全球化的不断深入以及信息化技术的升级和大数据时代的到来，新一轮的产业转移浪潮逐步形成，并且呈现出了一些新的趋势。

第一，在跨国公司的推动下，国际产业转移的规模不断扩大。比如到2007年，全球外国直接投资总额增加到了1.83万亿美元，达到了一个顶峰。直接投资规模扩大了，产业转移的脚步也会更快。换句话说，在国际分工中，技术的优势越来越明显。

第二，产业转移的方式更加多元化，所涉领域也更加广泛。比如，现在的产业转移已经和30年前的传统转移——劳动密集型、资本密集型、技术密集型的梯度性转移——不同，是把产业链的两端即研发、制造、销售、服务等价值链的各个增值环节进行转移，行业的区分界限越来越模糊，不同产业环节的整合度越来越高。尤其是服务业的跨国并购和重组增多，信息产业成为强国产业发展的重中之重。

中国的机会在哪儿？

当前，中国已经成为全球最大的工业国，我们的经济崛起就来自承接了西方的制造业转移，中国的崛起也是全球产业大转移过程的一部分。相比之下，西方主要发达国家却出现了产业空心化的现象。同时，随着大数据时代的到来，制造业又重新回到了产业结构的中心位置，成为了新的重要增长力。

☆突破：大数据是国之利器

大数据不但意味着一项数据分析技术，还代表着知识产权的升级。中国向产业链高端迈进，离不开大数据技术，也必须依托自己较为完善的工业体系和后发优势，发展自己的大数据产业，整合信息产业和制造

业，才能在国际分工博弈中后发制人，超越在前。

时间倒退二十年，从20世纪90年代开始，互联网技术发展让生产过程可以相当容易地委托给国界之外的代工企业，所以“离岸外包”模式迅速发展成全球性的巨大体系。那时中国的中小企业成为了全球离岸外包的最大承接商，承担了一个接受外包的角色。这就是链条的中下端。

如何突破这个链条，向上攀升呢？这要靠新的知识产权和技术。因为决定产业链中的利润分配的，不是“谁生产”，而是“谁掌握了标准”。标准的制定者同时也是技术的拥有者和授权方，它们是可以对标准本身进行“知识产权化”处理的，然后形成一套使它自己利益最大化的规则。

中国进一步发展需要突破的核心瓶颈——产业升级，实际上就是一个如何使中国自己的生产链条相互之间形成配套的问题。大数据技术无疑为中国提供了这样的一个契机，那就是占领新的高端，制定“中国规则”，并在新的规则下进行符合中国利益的产业分工。

CHAPTER 3 中国如何搭上大数据快车？

中国的大数据现状

中国大数据市场元年为2013年，因为在这一年，一些大数据产品已经推出，并且部分行业产生了大数据的应用案例。据估计，未来的四年将迎来大数据市场的飞速发展。

不得不说，数据已经和我们息息相关，不知不觉中已经渗透到每一个行业和业务职能领域。全球知名咨询公司麦肯锡在一项研究报告中指出，人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。大数据这一概念，不仅引起计算机行业的关注，也成为了其他行业内的一个重要概念。

☆信息现状

众所周知，数据是信息的载体。我们对于信息的理解不能仅仅局限于“数字”，而是存储在计算机里的一切信息，视频、音频。由于数据的不断增多，数据所承载的信息量也就越多，因此人类知识的边界会不断扩大，未来也许会发生“信息爆炸”，在海量的信息面前，人类必须学会利用数据，以及数据里的信息。

曾经有一个技术工作者和我说，他亲身经历了从小数据到大数据的时代转变。他的工作是数据库程序员。据他所说，当他在美国工作时，能感觉到数据背后有一股强大的力量。他还告诉我，我们进入大数据时代最好的证据就是数据库变成了数据仓库。然而当他回国之后才发现，原来我们的大数据处理的各个环节，和美国有相当大的差距。

大数据时代，是我们面临国际竞争时的又一次挑战。我国的传统文化与国外有相当大的差异。我们的传统是重观点、轻数据的。举个例子来说，我们国家的社会学科是偏重定性研究，而国外则相反，比较注重定量研究。因此，当我们面临着数据带给我们的机遇和挑战时，能不能从中找到创新的方式，正视传统的不足，这是至关重要的。

然而我们身处大数据时代并不是与绝对的安全画上等号。我们的网络安全需要政府部门制定专门的政策和法律来保障隐私。

隐私的边界在于你个人对自己的隐私必须要有控制权，也就是你可以任意地管理和删除自己的信息，并且不会被别人窃取自己的身份信息进而造成困扰。

我们国家在这一层面上，做的显然还不够。中国社会迈进大数据时代的制约和障碍，以及给予我们每个人在隐私层面的考虑，是每个人必须要思考的问题。

☆人的现状

与大数据时代相适应的，则是能够处理海量数据的IT技术人员。而

随着大数据产业的迅猛发展，IT技术人员的需求也与日剧增。其中对IT技术人员的需求不仅仅是数量上的需求，更重要的则是质量上的需求。

在2012年的大数据世界论坛上，有嘉宾呼吁：“与大数据相关岗位的人才短缺。欧美也在中国市场寻找这方面的人才，但是他们不知道中国这方面更匮乏。将来一个国家的竞争力很大程度上决定于分析人员，将来的决策都是通过数据来说话，通过数据分析得到结果来做决策。所以，分析人员的水平对于国家的竞争力、对于一个企业的竞争力来说是非常重要的。”

根据我们的调查，目前国内数据分析人员整体逊于国外分析人员。国内数据分析人员的薪酬水平也远低于国外的数据分析人员。在这样的市场竞争环境下，是很难出现高水平的数据分析人员的。

刚从高校毕业的陈同学有机会进入英特尔公司从事平台软件的研发工作。如果不是在大数据的浪潮之下，也许陈同学将会继续沿着工具型软件的研发道路走下去。

但是现在，各个企业（包括英特尔在内）都对大数据产业非常重视，因此也对陈同学的职业发展道路产生了非常巨大的影响。陈同学决定将大数据平台作为自身未来发展的一个方向。他说：“无论是在科学理论研究，还是在产品开发层面，大数据的热门和快速的技术更迭都能满足我个人对新技术兴趣爱好的需求。此外从职业发展层面，大数据也能带来更多机会。我看过一个报道，显示大数据相关的职位在IT技术领域包揽了高薪排行榜的前三名，这也是整个市场对大数据产业热度，以及大数据产业人才相对匮乏的一个反映，我相信在未来几年，能满足这个产业发展需要的人才将会得到更多、更好的发展空间。”

然而一个严峻的现实摆在我们面前，如何才能培养数据分析人才呢？

首先，在大学本科的课程中，更加重视对一线实际数据的分析能力的培养。在教学方面给予足够的重视，是非常非常重要的一环。因为大学本科阶段是打基础的重要阶段，因此若想培养数据分析人才，对数据分析能力的教学是至关重要的。在人才培养阶段，绝对不能让学生轻易“糊弄”过去。

其次，企业主应该重视数据分析在企业决策当中的支撑作用。大企业不应该止步于使用Excel这样的统计分析软件，而是应该使用更加专业和精确的分析软件。另外，可以从国外引进先进的分析方法和模型，定期对员工进行培训，这样可以逐步提高研究人员的业务水平。

另外，政府也应该制定一些针对高水平数据分析人员的鼓励政策。为了防止人才市场上“劣币驱逐良币”的不良竞争出现，政府也应该科学

地加以宏观调控。

在大数据时代，数据处理的能力相较几年前已经有了质的飞跃。但如何将这些数据在各行各业中加以利用，并且做出正确的决策，是数据分析人才的工作。

大数据虽然已经在中国遍地开花，但它能否硕果累累，还要取决于高水平的数据分析人员。

☆技术现状

大数据本身的特点，一定程度上也是其技术的必然要求。在处理海量数据的同时，对数据处理人才的需求以及数据处理的技术需求是同样重要的。而大数据技术的真正价值在于对未来发展作预判，因为它本身的优势就是分析和预测。而对大数据的技术应用可以是某个企业，也可以是某个产业。

大数据的技术重点在于应用。在未来，大数据的战略将以应用为中心，以效益为导向，并且将更多地用于政府和企业。

然而，在大数据处理的技术应用方面，国内外还是有很大的差距的。以美国为代表的发达国家已经开始把大数据的利用与大数据技术开发视为国家的一项战略性任务。我国对于大数据的技术应用还处于基本的起飞阶段。况且，我国作为一个发展中国家，经济处于起飞的阶段，非常容易急于求成，并且拥有浮躁的心态。若想在大数据的技术方面有非常重大的突破，必须在大量应用的基础上发展适用的技术，先学会如何把数据应用处理好。

我们国家在数据技术应用层面也有着特殊的国情，这一点我们不得不考虑到。好的一面是，国内大数据市场需求广阔、后续增长潜力大、投资前景好等等。然而与之相伴的另一面则是，人群庞大、复杂程度高。还要考虑到各种政策、理念和历史因素，因此综上所述，中国是世界上最复杂的大数据国家。

用宽带资本董事长田溯宁的话说："现代历史上的历次技术革命，中国均是学习者。而在这次云计算与大数据的新变革中，中国与世界的距离最小，在很多领域甚至还有着创新与领先的可能。"

大数据已成为21世纪人类可开发利用的重要资源。工业和信息化部通信发展司副司长陈家春对此表示："我国大数据产业同样也面临着人才匮乏、数据资源不够丰富、数据开放程度较低、相关的法律法规不完善等问题。"

面临着如此复杂的时代环境，国内的互联网公司必须在这波数据服务浪潮中迅速找准自己的位置，把碎片化的数据用种种手段整合起来并加以利用，增强自己掌握数据结构性，加大数据关系性。这将会成为未

来产业发展的一个显著走向。而这些数据则是在获取和整合更多的用户行为基础之上才能实现的。例如我们所熟知的云计算、各种木马、cookie之类的产品和服务等等。

虽然这些尝试仍旧处于基础阶段，但是一些硬件层面的变革正在悄悄发生。而硬件层面的技术对于数据本身的收集、存储和分析是不可或缺的。

国际IT企业都推出了针对大数据的产品，并嵌入自身的产品服务之中，如IBM、EMC、惠普等。国内IT企业也认识到了这个问题，开始在原有的产品基础上加大数据领域的研发和投入，并也有了初步的产品和方案。

具体来看，有两点非常有利于中国信息产业的跨越式发展。

第一，大数据技术以开发新资源为主，并且这项技术对任何一个国家都是开放的。到目前为止，尚未形成绝对技术垄断，中国公司同样可以分享蛋糕。在过去的几十年中，中国信息产业长期处在产业链的末端，赚取低端的利润，一直落后于国外。虽然国家大力扶持技术方面的创新，甚至会提出创新驱动发展战略，比如对CPU、操作系统、办公软件的创新，但是鲜有成功的案例。然而大数据处理是一个新兴的领域，中外公司几乎站在同一起跑线上。因此有业内人士表示，某种程度上，单纯考虑狭义的大数据处理技术，中外差距仅有5年左右。

第二，由于中国人口众多，经济增长迅速，因此决定了中国的数据资产规模一定冠于全球。数据资产规模巨大，客观上为大数据技术的发展提供了前提。然而如何能够有效地利用这么庞大的数据，则需要政府、学界、产业界、资本市场四方通力合作，这样才能确保最大程度地开放数据资产，保障国家数据安全，促进数据的关联应用，从而释放大数据的巨大价值。

事实上，我国目前对大数据资源的价值和应用，各地方普遍存在“数据割据”和“数据孤岛”的现象。政府、学界、产业界和资本市场尚待达成一致的认知。归根结底，除了技术上的欠缺，对大数据缺乏深刻的理解和认识，缺乏数据开放的理念，是阻碍我国大数据技术在各行业实现大发展的关键因素。

因此数据开放的意识和能力是最重要的一点。在数据公开方面，应该由代表公共利益的政府成为数据开放的潮流引领者以及政策制定者，而不能仅仅靠个人和企业。从过去的网络发展来看，未来应该是走向集成、动态、精细和主动的新阶段。

我国正处于现代化发展的新时期，联系到其中所面临的种种问题，要想实现中国制造到中国创造的改变以及缓解教育、交通、医疗保健等

各方面挑战，必须通过大数据这种创新方式来解决问题，创建新的产业群，为政府做出科学的预测。

大数据事实上是一条新的产业链，将其扩展开来，会形成规模庞大的基础产业。

虽然目前政府和一些产业积累了大量的数据资产，但是好多行业都缺乏行之有效的算法来充分挖掘大数据背后的真正价值所在。

大数据产业链包括：数据技术产业、数据采集业、数据加工业、数据应用业。数据处理的每一个部分都可以成为相当规模的产业，关键就在于如何实现信息化的转型，这关乎到信息全部领域的生态链建设，对于大数据产业的发展也是至关重要的。

在大数据落地应用方面，国内厂商也有自己的创新。例如神州数码、东软等IT软硬件企业已经看到了大数据的力量，并且开始在原有的业务和产品基础上加大数据领域的研发和投入。浪潮基于对数据产业的认知与积累，采用新型技术体系架构，推出云海大数据一体机解决方案。

虽然我国在数据处理和应用方面和国外有明显的差距，但是以未来的视角看，只要无论是政府、互联网公司、IT企业还是行业用户，都以开放的心态、创新的勇气拥抱大数据，大数据时代就一定有属于中国的机会。

☆文化现状

大数据体现出来的首先是一种世界观，其背后是我们的文化思维所表现出来的力量，以及看待这个新时代的方式。

大数据专家维克多·迈尔·舍恩伯格在其相关书中解释，云计算在获取海量数据的同时，也带来了数据的混杂性，这给传统的数据分析带来了困扰。

以往，我们习惯于由数据得出具体结果。然而，大数据时代，应当关注的是相关关系，而不是因果关系。大数据对于整体数据的获取带来思维方式的改变。它促使我们更加注重事物与事物之间的联系。

大数据带来的另一重要改变是：更多事物可以数据化。例如，社会热点的走向可以数据化，购物习惯可以数据化，社交关系可以数据化……未来我们的生活将会时时刻刻存在着对于各种大数据的分析，这也要求我们每个人都要有分析数据的能力。

有很多例子能够看到我们身处的大数据时代可以走得更远。

我们分析数据的时候，分析的样本由部分数据变成了总体数据，云计算能够帮我们搞定所有想要的分析结果；如果你上淘宝，登录支付宝账户，点开电子对账单，你就能看到自己一年的消费曲线图；当你在当

当网下单某本书后，它会提醒购买这本书的人中有30%也购买了另外一本书……所有的一切都是基于对大数据的分析。

这些数据可以导出商业潜能，更能导出社会走向。这样庞大的数据分析，在小数据时代根本无法做到。

你可能要问，大数据真的能获得最全面的信息？一些企业已经通过大数据的分析得到了很多的数据分析结果，而这些结果在我们看来是很匪夷所思的。从这些数据中你可能发现，天蝎座的男性在2012年平均消费额最高，上海人用支付宝缴纳水电费的频率最高。7-11便利店通过分析零售终端的数据，得出了“温度低于15摄氏度，暖宝宝的销售量便增加5%”的相关关系。于是，只要温度低于这一度数，店内的暖宝宝就会上架。

当然，正如一枚硬币有正反两面，任何事物都是一把双刃剑。我们身处大数据时代，不得不时时刻刻被它敏锐地监控着我们的生活。仔细思考一下，亚马逊、当当网、淘宝似乎都在监视着我们的购物习惯，百度、谷歌似乎在监视着我们浏览网页的习惯，微博、人人网、朋友网似乎对我们以及我们朋友的关系无所不知。

如何才能够让大数据不侵入我们的隐私边界，是需要我们每一个人严肃思考的问题。

☆企业现状

在大数据的浪潮已成翻天覆地之势时，在世界性的数据革命已不可逆转之时，中国的企业和企业家们做好准备了吗？

马云在发表自己的卸任演讲时警告中国的企业家：“大家还没搞清楚PC时代的时候，移动互联网来了；还没搞清楚移动互联网的时候，大数据时代来了。”

和我写这本书的目的一样，马云在提醒中国的大众和决策精英：大数据时代真的已经来了。这不是什么“狼来了”的不入流故事，而是一个事实。

问题已经抛给了我们的企业。

Web 2.0时代是以社交网络的兴起为基础的，因此大数据应该以人的关系为基础，通过信息的生产、交换，从而产生信息的巨大爆炸。对于中国的企业来说，最重要的是去改变企业流程与文化，而不是怎么来做选择、来实施大数据分析。要知道，“第一个吃螃蟹”的企业将会保持领先，而跟随者将会错失机会。因此数据将是下一个大的自然资源，将会区分每个行业的胜者与输家。

随着大数据时代的到来，企业应该在内部培养三种能力：整合数据的能力、探索数据背后价值的能力、快速实时行动的能力。正如某IT行

业的精英所说："如果企业在信息治理上培养出这三种能力，对未来大数据时代的驾驭能力会增强，面临到的挑战会降到最低。"

本质上说，基于大数据的处理和分析才能为企业带来巨大的增值价值，而大数据本身并没有太多价值。

落后是现实——中国企业需要更多的参与度

"国内IT尤其是软件企业在布局大数据方面，已经落后。"某IT行业高管认为，这主要是因为国内企业在数据库、数据仓库、商业智能等领域基础薄弱。

"国内企业在这方面仍有机会，但需要找准行业与切入点。我比较看好互联网公司，像百度、阿里巴巴、腾讯，这样的互联网公司比传统的IT和软件公司更有机会。"

实际上，尽管作为中国最大的电子商务公司，阿里巴巴已经在利用大数据技术提供具体服务，但是国内没有一家真正朝大数据方向努力的公司。

例如，阿里巴巴通过掌握的企业交易数据，借助大数据技术自动分析判定是否给予企业贷款，全程不会出现人工干预。截至目前，阿里巴巴已经放贷300多亿元。

不过，某专业人士并不认为这些是真正的大数据。他表示："国内很多公司都在做分拆，并且产品数据也不相通。短期来看，这样可以提高效率，但是长期看，这是反大数据方向的。" 未来能够称得上大数据的公司将是Facebook、苹果、谷歌等这样的平台型公司。

蛋糕的划分——互联网公司走在前面

随着我国网民数量的增加以及第三产业的迅速发展，我国互联网行业取得了令人瞩目的成绩，在用户行为积累和数据处理方面积攒了一些经验，形成了覆盖数亿网民、辐射各行各业的全产业链。

然而就整个产业而言，与国外互联网行业的发展相比，我国的互联网行业仍处于大数据时代的萌芽初期，这也是我们所要面对的现实之一。与传统的几大行业相比，我国只有搜索引擎和电子商务才能在商业模式和经营水平的层面上对大数据进行有效利用。传统行业比如电信业、金融业等等都没有能够开展有效的应用。

但是从另一个角度来说，互联网公司本身的业务和行业特征，能够在数据的分析上做得相对领先一些，更有可能直接形成突破。具体来看，国内这方面最有发言权的莫过于百度、新浪、阿里巴巴、腾讯等几大互联网巨头。从这几大巨头的大数据应用和发展的规律中，我们能够看到大数据应用的一些前景。

百度对于大数据的应用莫过于在搜索的基础上，它推出的百度指

数、百度数据、百度风云榜等等，都通过广大网民的搜索行为数据为各行业提供网络搜索咨询报告，或者为广告商提供相关的搜索统计数据来赚取利润。百度的最大优势就在于其庞大的用户群和用户黏性。但是问题也随之产生，搜索结果个人化，搜索结果与广告之间的相关度上进展有限，百度的大数据应用是否能够走得够远，我们拭目以待。

有种说法，“电商行业的现金收入源自数据”。阿里巴巴集团旗下的淘宝、天猫等电子商务网站，在大数据的应用上又有哪些创新和发展呢？

它曾经研发出创下“巨大声誉”的阿里询盘指数，即买家在采购商品前，会比较多家供应商的产品，反映到阿里巴巴网站统计数据中。阿里巴巴前董事长马云曾在2008年初通过观察到询盘指数异乎寻常的下降，为企业做出了科学的预测，从而帮助了成千上万的中小制造商。商家可以通过这一服务及时了解本行业内的各种情况，并且做出科学合理的经营决策。

腾讯作为中国互联网另一个当之无愧的巨头，又是如何在即时通讯工具如QQ、微信上来实现大数据的应用呢？

事实上，腾讯在数据领域的布局主要集中在产品上，但是目前能够对外开放或者深度加工打造出的单独服务还不多。有两款应用分析工具值得注意：腾讯分析和腾讯罗盘。

还有其他互联网公司的大数据应用模式值得我们观察：携程网在用户习惯深度挖掘基础上形成了一套新型服务模式和服务体系；凡客在2011年成立了数据中心，研究新产品的上架与新用户增长的关系，希望能够实现互联网的系统化和数字化的管理；新浪微博则成立了数据部，初步形成了一个小型社交媒体数据分析挖掘生态……无数互联网企业开始重新审视自己的行业定位，在大数据时代的背景下，将数据资源和数据价值提升到自己的核心战略中。

综上所述，大数据的出现既为互联网公司带来了机遇，又为新兴的创业公司提供了巨大机会。

数据量的增加是每一个企业所面临的机遇和挑战，如果能够有效地利用，就可以进一步地去探索如何才能把个性化和智能化相结合，从而实现广告业务的增加，创新一种全新的商业模式，寻找到增加用户黏性的结合点，有效开发新的产品和新的服务，从而降低运营成本，最好能够通过大数据来实现规模效益和经济效益，以大数据产业为纽带，进一步带动我国经济的发展，促进数据流通的合理循环。

机遇永远与挑战并存。对于每一个企业来说，这就像冬日的朝阳，迎着朝阳前进，必然会看到曙光，但也要做好保暖的措施，找到那个合

适的支点，你才能够撬得动地球。

追赶者的中国——走到至关重要的十字路口

迈克尔·坎特作为美国贸易代表参加了中美入世谈判，后来又担任克林顿政府的商务部长。他数次参与中国的入世谈判，经历了中国入世过程中的关键环节。

对中国的发展，迈克尔说："我不是什么预言家，我也无法预测。不过，我想中国将永远是世界的一个强国。这是一个好消息。中国使那么多人口脱贫，超过了历史上任何一个国家。中国和美国关系至关紧要，而且需要更加紧密。中国经济将在不久的将来超过美国，成为世界第一大经济体。中国目前正在转变增长模式，提高在全球价值链中的地位。"

在最艰苦的阶段，中国科研领域里许多成绩出色的人离开了祖国，去国外发展，他们觉得中国没有希望了。特别是那些学航天、宇航、飞机制造、现代船舶、发动机等学科的高科技人才，他们中的许多人都选择了离开这个国家。

虽然他们多数人去国外后只能在一些小公司打零工或者当副手，但他们仍然义无反顾地去到了欧美社会。当时的中国没有任何平台给他们施展技术才华，而在国外一个月就能赚到几千元的薪水，这几乎等于国内研究机构干一辈子的全部工资了。

那些当初选择留下来和中国一起成长的人们呢？他们可能不是成绩最好的那批人，没有达到让欧美强国收留他们的标准，但就是这些人，坚定地与自己的祖国站在了一起。事实证明，这些留下来的人和国家一起奋斗，反而成就了一番伟大的事业，选择离开的那批人，则多数在默默无闻中浪费了自己的生命和才华。

这些年，中国经济飞速发展，科技也在快速进步。比如航天、军工、电脑芯片、数控设备、精密加工和卫星导航等国之重器不断研制成功，创造了它们的人也将名留青史。他们牺牲巨大，举世罕见。中国人将在这条充满荆棘的道路上书写他们的故事，为追赶者的中国付出一代又一代人的心血。

和30年前相比，中国已发生了巨大变化，近期的发展目标不再是解决温饱，而是实现全民小康，超越中等发达国家，进而实现民族复兴。这是一个已经持续了一百多年的目标，从近代起中华民族在科技领域就一直在苦苦追赶，新中国成立60多年来，中国正以令人难以置信的速度迅速缩短着与国际最高科技水平的差距。奋力追赶，这是一次全国总动员，是一次全民族的超越战略。在历史上屡次与全球科技革命失之交臂的中国，现在正急迫地呼唤着科技创新。

未来的“中国制造”所面对的，将不再是仅靠价廉即可取胜的低端市场。但是，随着这一进程的加快，中国企业无论是在国内还是国外，都日益感受到了跨国巨头带来的竞争压力。

中国现在走到了一个至关重要的十字路口，我们必须进行产业升级，才能躲开“陷阱”。

更重要的是，随着科学发展观深入人心，中国人越发深刻地认识到，以前单纯依靠加大资本投入、人力投入和物质消耗获得经济发展的模式，不仅面临无法逾越的增长极限，而且过度“透支”将给我们的未来带来深重的无法承受的伤害。而且，这些年来，由金融危机引发的全球经济衰退，也为我们必须加快转变经济发展方式、推动产业结构优化升级提供了更加直观的例证。

现在，中国面临着新的挑战。

☆产业升级

中国未来的产业布局如何进行，这是摆在中国政府以及有使命感的企业家面前的大问题，也是一项大挑战。为什么必须进行产业升级？因为中国仍然正在走向工业化，还没有真正地完成工业文明的建设（包括信息工业和新科技革命），中国仍是后进学生。这是一段很长的历史时期，因此中国的产业升级已到了关键时刻，也到了新一轮的深层次阶段。

第一，要实现从中国制造到中国创造的转化，而且是要由中国制造走向有更多中国创造元素的中国制造，因为创造是一个基础的产业，制造则是这个产业的基础，两者是统一的而非对立的。

第二，实现产业升级的同时，也要做好产业转型。我们要建立一个强大的工业国家和强大的制造业，让东部和中西部地区平衡发展，而不是把已经拥有的丢弃。正确的做法是将东部产业向中西部迁移，从而为东部地区的产业升级腾出空间。

第三，控制虚拟经济与实体经济的比例，达到良性的平衡。我们要让经济建立在坚固的石头上，而不是沙滩上。石头是制造业，沙滩就是金融等虚拟经济。现在，就连美国这样强大的经济体，也在寻求复兴制造业，促进制造业回流美国，我们难道为了产业升级就可以付出制造业空心化的代价吗？当然不能。

因此，中国在产业升级的时候必须强调工业强国，坚持以制造业为经济基础。在这一点上，大数据有着至关重要的意义，因为它对于一个国家的工业实力有着几何倍级的增升效能，我们完全可以通过大数据技术来提升产能，节约原材料，并且帮助产业升级正常进行。

☆人口数量太多和人口结构不平衡

人口红利是一个描述短期利益的经济学术语，中国经济的飞速发展，离不开人口红利的贡献。

这个学术用词也有它令人振奋的积极一面。但是，计算一个民族和国家的长远利益，却需要综合考虑。也就是说，人口过多仍然是中国人口的首要问题，人口结构的失衡也更令我们感到担忧。

随着中国的产业结构升级，我们对劳动力密集型产业的需求逐渐减少，对于劳动力的需求也将从数量转向质量。那么，科学地控制人口就成为了政府的人口工作要面临的任务。在这方面，大数据技术有充足的舞台可以施展。

☆信息化升级

信息文明也被称为“后工业文明”，它是工业文明的高级阶段。在这一阶段，人将成为技术和信息的主人，使用信息通讯技术来驱动经济社会转型，在产业集聚、工业生产、商业零售、社会生活等各领域都产生全方位的革命性的变化。

中国也正面临信息化升级的重大使命，信息化技术的实现，将导致整个中国工业产业和商业形态发生积极变化，进而使得整个中国经济社会都产生革命性的变革，帮助中国顺利地完成产业升级，向发达国家迈进。

根据预测，到2020年，全球将有500亿个无线设备联入互联网，数量是今天的10倍。全球的无线联网设备将从目前的手机、汽车，延伸到医疗、教育、能源、制造业等领域，一直到全部的终端目标。但是中国还有许多工作要做，比如，中国目前的宽带网速不到韩国的1/10，价格却比国外高出很多，怎样实现宽带投资的提升，加速中国的信息化进程？云计算在世界范围内已蓬勃发展，中国也有相当多的企业在加速布局，并取得了一定的进展，它们应如何更上一层楼，发展出中国的成熟的云计算产业，为大数据时代在中国的开启铺一条宽阔坚固的光明大道？这些都是具体的问题，也都非常重要。

我们的传统产业也具有巨大的升级潜力，信息技术和产业相结合的系统集成，可以对中国的传统产业实施改造，然后实现高附加值的增长，在另一个领域帮助中国的产业进行整体的升级。在新的信息技术推动全球经济发生急剧变革的情况下，如何从全局意义上把握这种形态，实现中国占据全球经济制高点，成为了一项迫切和重大的课题。

☆社会管理模式升级

社会管理模式是指一个国家的管理体系。我们知道，战争的成效取决于作战的模式，社会管理的成效当然也就取决于社会管理的模式。在任何领域内，“模式”都是科技含量高度集中的地方，是“能量”的重中之

重。大数据在社会管理中最大的价值体现，就是对模式的完全创新，它颠覆了人们的习惯，并创造了一种新的高度有效的模式。

不愿放弃过去已经熟悉的做法，这就是习惯的力量。管理就是如此，人们总是不喜欢在管理上进行创新。守旧或有本能的一面，不管是国家、企业还是个人。

耶鲁大学的教授金·豪尔说："什么是现代化？现代化就是有系统地、持续不断地、有目标地运用人类的各种能力控制自然和社会环境，以达到人类的各种目的。"我的解读就是，现代化需要不断地创新，提升社会治理模式，这是人类文明进步的基础。

要提高中国的社会治理水平，就需要改革和更新当前的治理模式。就像在战场上，军队的作战模式随着武器的发展不断进步一样，大数据技术的发展，也使现在的社会治理模式迎来了一场可能是人类历史上最根本性的变革。比如可以在福利、行政、交通、应对突发灾害、维稳等工作范围中使大数据技术发挥更大的作用。

如果说西方国家由于它们雄厚的积累和先发优势，即便国家需要突破某些障碍，也只是面临一两个问题的话，那么今天的中国就是在面临数个"十字路口"，而且不容许犯下任何一丝错误。

消除“数据割据”与“数据孤岛”

如果我们把数据比喻成某种资产，在割据状态下，数据市场呈现出来的是这种形态：每个人都躲在自己的房子里闭门造车，收集和增加数据（资产），捂紧口袋，待价而沽。互相之间获得数据是非常困难和缓慢的，中间需要讨价还价、建立信任和足够多的时间成本，而每个数据商（独立的家）就像一座座孤立的小岛。

当我们尝试数据的治理进入到实质阶段时，就会发现有这三个问题在阻碍我们的工作：数据割据、数据孤岛和数据质量。它们既是统一的整体（危害通常一起爆发），又可以在某些阶段单独呈现。

数据割据——因为制度、部门保护主义或小团队利益等人为的因素造成的数据分散的现象；

数据孤岛——因为技术差距和遗留问题等形成的数据分散与无法集中共联的现象；

数据质量——主要包括数据的真实性、完整性和一致性。我们都知道数据质量的好坏直接影响着“数据资产”的价值，但解决却非一蹴而就，需要各方面的综合提升，比如技术、制度、文化等多领域的努力。

数据割据现象更多存在于国家各部门、各地方之间，大型企业也会造成数据割据现象。基于它必将产生的、对于效率的危害，数据割据是我们首先要消除的不良现象，它违背了大数据时代的精神，急需人们拿出解决办法和协作诚意。

具体来说，割据造成的数据孤岛有几个让现代人已经无法忍受的特征：

一、数据使用者（用户）的成本不断叠加，在一项服务中重复支付成本；

二、阻断技术的进步，难以实现产业联合；

三、“跨岛”合作的环节烦琐缓慢，信息共享缺乏实时性。

因此，虽然国内的各行各业都对数据资产充满了向往，将手头的巨量数据转化成盈利，这是一个光明的前景，但人们更担心的则是“数据割据”的现实，它让这个前景蒙上了一层厚厚的阴影。

比如，淘宝网对八度网络公司的警告，认为它的“超级数据平台”侵犯了淘宝“数据魔方”的软件著作权。中国政府没有办法进行表态，因为我国目前还缺少数据资产归属权、使用权的立法，也没有界定网络上公开的商品价格的数据，它是应该只属于淘宝公司呢，还是应该打开大门，让其他公司也有权利分享。淘宝和腾讯微博先后屏蔽掉百度的蜘蛛爬虫也是一个明证。

现在国内的互联网巨头都掌握着大量的也是部分的消费者数据，像百度、腾讯、阿里巴巴集团三大互联网公司分别掌握搜索、社交和消费数据。这三方数据如果能汇总在一起，可以拼凑出一个完整的网上信息的数据关联图谱。但现实是它们之间没有合作的意图，相反却是高筑墙，阻止数据外流，至少不会让对方轻松地得到。

从大数据的本质来看，其开放与分享精神在中国碰上了数据割据的壁垒。人们知道只有分享才能充分发展大数据的巨大价值，实际做起来却不是这么回事，理想和现实之间还存在着几乎无法跨越的鸿沟。

一位供职于国内某知名IT公司的大数据专家梁先生说："中国的互联网巨头本质上都还是封建割据的思维，没有认识到信息的流动才是互联网的精髓。"梁先生长期关注大数据，他认为，这些互联网公司都认识到了数据将是未来发展的战略性资产，所以它们不会轻易拱手让人。但正因如此，才造成了中国目前的大数据产业缺乏完整性和一致性，使得可利用价值大打折扣。

特别是对处于大数据产业中下游的中小型企业来说，相对于淘宝、百度等巨头，它们没有技术优势，如果再缺乏数据源头的支持，将真正地在运营和布局上处于尴尬的境地：虽然有锅，但是无米。

在业内进行调查时，很多经理人都反映说，即便巨头们口头上承诺会开放，会让其他公司分享数据，可实际上是"挂羊头卖狗肉"，没有真心诚意将数据源开放，或者它们只允许在其各自的平台上运行。

在政府部门中也存在这个问题，比如我们社会中的个体信息，封闭在工商部门、银行、保险、公安、医院、社保、运营商等不同的机构手中，融合起来非常困难。基于部门利益保护主义，它们对信息共享缺乏动力，这是中国必须尽快革除的障碍。只有打破孤岛，我们才能看到中国的大数据时代的曙光。

大数据并不是存在于某一个部门之中，它发挥作用也不是某一个单独的部门可以实现的。政府需要从横向和纵向同时比较一些数据，来得出最贴近于事实的结论。因此，解决数据的割据和信息孤岛问题、提升系统建设的能力和规划势在必行。从技术层面看，云计算的高度灵活性正好对应了中国政府的需求。

我的建议是，大数据产业的发展，应由各级政府带头实现等级制数据开放共享。从上到下制定一系列针对性的政策和法规，引导和推动各行各业来开放数据，进行数据合作，将所有的数据熔到一个炉子里。在把握巨大商机的同时，也要做好利益的分配，并注重保护特殊信息。

工信部的规划：四项关键技术创新工程

较早前，中国就把云计算列为了新一代信息技术产业的重点领域，而且将在“十二五”期间给予大力扶持。在调查中，工信部的总经济师周子学介绍说，具体措施包括加快云计算技术研发的产业化，组织开展云计算应用试点示范，着力完善产业发展环境等。

中国已经开始行动，迈出了实际性的步伐。在“十二五”时期，工信部将把加快云计算产业列为重点，推动其服务的产业化，提高创新和信息服务的能力，并且要推进核心计算的研发，最终形成云计算系统解决方案的提供能力，使重大的产品产业化，形成服务标准和规范的管理体系。

在“十二五”规划中，工信部列举了与大数据有关的四项关键技术创新工程。它们是：

1.数据挖掘：解决对数据收集的能力；

2.海量数据存储：解决对数据存储的能力；

3.信息处理技术：解决对数据分类与分析的能力；

4.图像视频智能分析：解决对数据的高级分析的能力。

这些都是中国大数据的重要组成部分。对此工信部的官员还说：“中国未来将出台更多的优惠政策，扩及中小型的数据中心，使大数据发展的政策空间更加广阔，利于企业的参与。”

中国规划建设的超大型数据中心的规模为1万个标准机架、拥有超过10万台服务器。从2011年以来，中国规划和建设的超大型数据中心达到了20～30个，它们分布在全国各地，也从国家获得了优惠政策。将来，这个优惠范围还将扩大到中小型的数据中心，逐步实现向基层发展。

而且，为了避免盲目地建设数据中心，在2013年初，中国多个部门就联合发文，共同发布了《关于数据中心建设布局的指导意见》，目的就是通过加强顶层设计和规划，引导数据中心的合理布局，让中国的大数据产业健康发展，走出一条富有中国特色的道路，为企业创造一个完善的政策环境。

建立大数据政府

“政府的前景？”在一次采访中，一位国外研究大数据趋势的专家对我说，“对政府而言，大数据更像是一种战略资源，它和能源一样，甚至比石油和天燃气的价值还要大。”随着互联网技术的不断发展，数据已经体现它的价值了，人们既创造数据，又活在数据的包围之中。整个世界不正是如此吗？中国也在行动，虽稍显落后，但已开始努力追赶。

怎样利用数据资源来提升政府的工作效率，利用数据技术来促进行政创新和信息的高效管理，建立真正的大数据政府，成为主要国家都在重点研究的一门高端学科。我们不但要有高产出的大数据产业，也要有高效率的大数据政府。而且，随着技术的不断成熟，大数据技术也将必然成为全球范围内国家治理的重要工具。

☆数据政府的机遇与挑战

奥巴马政府已经宣布了美国的“大数据的研究和发展计划”。在这项庞大的计划中，联邦政府的六个部门和机构宣布了超过2亿美元的投资，提高从大量的数字数据中访问、组织、收集发现信息的工具和技术水平。这是除情报领域的大数据变革外，美国政府在行政领域的重要布局，它们的目标是继新科技革命后再一次在大数据革命的浪潮中领先全球，成为其他国家学习的榜样。

美国一些相关的公司也行动起来，它们已经赞助大数据相关的比赛，并且给大学提供这方面的研究资助。常春藤大学也纷纷开设了一门全新的研究型课程，为美国社会培养下一代的“数据科学家”。

白宫发言人说：“我们还非常有兴趣支持建立一个跟大数据相关的论坛，这包括最新的公私组织之间的合作。比如欢迎非营利性机构来对公益性的服务数据进行采取、分析和可视化处理。”

中国政府也在有所作为。早在2006年，国家统计局就成立了社情和民意调查中心，首次真正地投入人力、财力来主动地收集民意数据进行分析，然后做出反馈和政策方面的完善。同一年，卫生部又制定了第一套最小标准数据集，用来向下级部门和各类社会组织收集相关的业务数据。

到2010年的时候，铁道部及交通部开始应用物联网的技术，通过自动采集收集环境数据为老百姓的日常生活提供相应的环境指标和预测。这堪称一次跨时代的进步，也正式宣布中国开始在基层拉开了大数据技术实用化的序幕。

这些年来，中国在民意数据、业务数据和环境数据方面的收集工作都在进行，无论是相应政策制定还是技术手段都得到了突破。我国利用

真实的数据来了解民意和老百姓的生活水平，对建立高效的服务型政府意义是重大的。因为不管是提升数据的收集技术，扩展更丰富的手段，还是向我们的国民提供更优质的电子政务渠道，都在对建设高效率的大数据政府产生积极的推动作用。

☆中国起步相对晚，急需提升系统建设的能力

我们既要看到进步，也要看到落后于人的现实，就像一位专家对我说的："中国还是起步太晚了。"这几年来，人们的生活水平急速提升，民众对于政府所提供的服务也有了更多和更高的新需求，例如个人业务异地查询办理、房产信息联网等。这些都需要继续推进，提升相关系统建设的能力。

从国民的需求可以看到，人们越来越希望政府可以即时、灵活、有针对性地提供相应的服务，这对于国内的电子政务建设来说，意味着不能再是简单的电脑化。西方主要发达国家早在20年前就开始系统地收集相应数据，作为制定相关政策的依据，推广和建设数据联网，我国政府对于数据应用起步相对较慢，反应较为迟钝。

在前11个五年规划中，中国的信息化建设依然更多的是围绕基础设施层面。中国人一头埋在基建领域，已专注太多年的基础建设了，即便在数据领域也是如此，比如"金税工程"，也是在弥补基层工作的缺失。这一方面在奠定中国政府提供服务的IT基础，另一方面也面临着社会更多需求的挑战。

虽然有一个省份的某地区在前年实现了社保卡的一卡通功能，也受到了全社会对于推广的期待。但接下来的进展并不顺利，因为数据的统一管理对中国的各机构来说一直是个很大的难题。在中国的整体电子政务建设中，各个机构之间就存在着许多信息孤岛和数据割据的现象，这些数据只有得到互通和有效的利用，才可以更好地创新一些服务。

在革除这些弊病的过程中，中国一方面需要从政策的制定和引导上进行改善，另一方面也应该利用更多技术手段进行改变，两手都要抓，两手都要硬。

☆利用数据说话的政府：从学习到使用

在IT界向人们强调大数据时代将来临时，无论是企业的信息化建设，还是政府的大数据系统的构建，都需要学习，并且不断完善，而且可能是一个漫长的过程。

从中国近些年的行动来看，我们可以明显看到政府已渐渐习惯利用数据说话，并利用数据为民众服务，治理和服务社会，这也已经成为了中国大力构建服务政府的重点部分。

随着收集数据的日益增多，政府也在深入地思考：是让这些数据只

是占用了大量的存储设备却毫无价值，还是成为政府制定相关服务政策和条款的依据？当然要选择后者，没有谁愿意费尽气力收集和保存一堆无用的数据。当各个部门的利益相关方都明白这样做对自己是有利的以后，政府的大数据时代才能真正到来。

在大数据的利用上，中国可以向一些先进的企业借鉴，看看它们是怎么进行应用的。另外从数据的数量上来看，我们政府的数据量相对于诸如制造业等传统领域还是太少了，所以在大数据的应用过程中，我们应该更多地学习如何使先进的经验和技术手段应用在传统行业中，帮助制造业获得突破。

最后在制度上，政府也应做出必要的突破。比如可以向企业学习，做到由一个部门来牵头，以保证数据的完整性和它最大限度的真实性。

CHAPTER 4 中国首批重视大数据的千亿公司

原动力：对信息共享的需求

任何一种新生事物，如果没有需求，它不会存在太久；如果没有大量的需求，它不会发展起来，也难以壮大。在我们身边，人们对大数据的需求是真实存在的，也是越来越大的，这就推动了相关产业的发展，给大企业搭上大数据的快车创造了条件。

大数据是一场席卷世界每个角落的深度变革，它不仅影响了人们的生活、工作，还影响到了人们思考问题的方式，最终对社会、政府和企业的组织模式产生了决定性影响。在五年前，还有人不屑地认为，大数据的主要作用不过就是帮助厂商更准确地了解消费者的消费行为，但五年后他们改变了当初这种幼稚的看法，因为大数据的功用远不止这些，它将在很大程度上影响人们的决策和行为模式。

需求决定市场，有多大的需求，就有多大规模的市场。

美国大数据工程师帕林海姆表示，很多客户现在仍然不明白大数据到底能给他们带来什么样的价值，虽然他们内在的需求十分强烈，但他们往往不清楚这意味着什么。不过，无论人们现在是不是理解大数据的内涵，绝大多数人面对大数据不会袖手旁观。人们总会自觉地去做那些可以满足自己需求的事情，所以现在越来越多的企业内部已经至少会有一个部门或一个人在做与大数据相关的事，来提升企业的价值，扩充市场并为自己的用户服务。

当云计算刚刚兴起时，也有很多人热衷于讨论这个问题："请问，云计算是不是一种具有变革性的创新技术呢？"然后有人回答："哦，我不看好它，相信我，过几年你就会看到它的消失，无人理会。"可是几年过去，人人都在用云计算。

那么，大数据是否也存在着类似问题呢？

云计算改变的是IT的消费模式，而大数据则在更深的层次根本性地改变了我们工作、生活和思考问题的方式。因为对于人们庞大的需求而言，大数据不仅仅代表了数据的量大，而且还意味着三种新的趋势：

一、海量的数据，让人们改变了看待事物与数据的角度，角度改变，当然结论也就不同了；

二、云计算的存在使人们有能力存储更大规模的数据，而且有更强的数据处理能力；

三、随着知识与技术的不断积累，越来越多的人能够进行大数据分析了。

这符合一项强大的新生事物的发展规律。并且，人们对于数据实时处理的需求正变得越来越迫切，我们每个人对数据实时分析的关注度已

经超越了对数据本身准确性的关注度，这正是数据量激增给人们的观念带来的巨大变化。

● 国家的需求：大数据共享平台

● 社会的需求：全民共享信息化

这个共享，不只是指对于海量信息的共享，更强调的是我们对于信息的筛选、处理技术的共享。我国在发展信息化方面，应该积极抓住大数据的发展契机，并及时将大数据上升到国家战略层面，来打造属于中国人的先进的“数据中国”。

在大数据驱动下的信息化，它的本质就是人们对于信息和知识的共享。没有人能拒绝这个大势，未来它一定会实现，就像工业革命的车轮碾碎了挡在路上的一切顽固势力一样。数据作为新一轮信息战的主角，它关系到政府、企业和社会等诸多的方面，也关乎我们自己。

云计算和大数据

提到云计算（Cloud Computing），你可能会问："它到底是一种什么概念呢？"

和其他概念上的计算一样，它是互联网的一种计算方式，在此基础上实现软硬件资源和信息的共享，而其提供的网络资源通常是虚拟化的，具有动态易扩展的特点。这种网络应用模式主要是基于互联网的相关服务的增加、使用和交付，最早是由谷歌提出的。

其实，"云"是互联网的一种比喻说法，人们把数据的计算方式形象化了。在过去，云被用来表示电信网；后来，"云"也被用作表示互联网和底层基础设施的抽象概念。狭义上的云计算是IT基础设施的交付和使用模式，指通过网络以按需、易扩展的方式获得所需资源；而广义上的云计算，则是指服务的交付和使用模式，通过网络以按需、易扩展的方式获得所需服务。

这既可以是IT和软件、互联网相关的服务，也可以是其他领域的应用。而且，这也就意味着计算也可以作为一种商品通过互联网进行流通。

在美国，美国国家标准与技术研究院（NIST）对云计算的定义是：云计算是一种按使用量付费的模式，这种模式提供可用的、便捷的、按需的网络访问，进入可配置的计算资源共享池（资源包括网络、服务器、存储、应用软件、服务），这些资源能够被快速提供，只需投入很少的管理工作，或与服务供应商进行很少的交互。

"云计算"的概念一经提出，就被大量地运用到生产环境中。国内的"阿里云"与云谷公司的XenSystem，以及在国外已经非常成熟的Intel和IBM，"云计算"的应用服务范围正逐日扩大。可以这样说，在大数据时代，云计算在未来的影响是不可估量的。

通常，"云计算"常常与网格计算、效用计算、自主计算的概念相混淆。下面，我们针对这几种计算方式的概念做一个基本的解释和区分。

网格计算：属于分布式计算的一种，由一群松散耦合的计算机组成的一个超级虚拟计算机，常用来执行一些大型任务。

效用计算：IT资源的一种打包和计费方式，比如按照计算、存储分别计量费用，像传统的电力等公共设施一样。

自主计算：具有自我管理功能的计算机系统。

事实上，许多云计算技术部署依赖于计算机集群也吸收了自主计算和效用计算的特点。但需要指出的是，云计算与网格的组成、体系结构、目的、工作方式等都大相径庭。

☆云计算的特点及应用前景

存储数据安全可靠

在过去，人们传统的数据存储方式是存进硬盘，很多用户苦于被病毒、木马攻击，或者由于一些疏忽操作导致数据的丢失。而云计算相比于传统的数据存储方式，给我们提供了一个最为安全可靠的数据存储中心，因为用户把数据上传到“云”以后，不用再担心数据丢失、病毒入侵的麻烦了。电脑出现故障甚至完全毁坏，仍然可以在另一台电脑通过“云”把数据复原。

但这种存储方式并不为所有人接受，很多人觉得数据存储应该和银行卡存钱一样，只有保存在自己看得见的地方心里才踏实，所以会以为自己的电脑才最安全。但事实又是什么呢？正如你所看到的，你的电脑会因损坏或者被攻击而遭到破坏，有些不法分子也会用各种手段窃取你的电脑数据。

前几年香港轰动一时的“艳照门”事件就是一个最典型的案例，受害者因为电脑送去维修而造成个人的数据外泄。如此看来，存在电脑硬盘上的方式未必安全。但如果文档保存在类似Google Docs的网络服务上，比如自己的私密照片、视频等，你就再也不用担心数据的丢失或损坏给自己带来意想不到的麻烦。在“云”的另一端，有专业的团队会管理你的信息，为你保存数据。而且，你不用担心数据泄露的问题，由于云存储有严格的权限管理，你可以放心地与你指定的人共享数据。

云计算对于客户端的要求低

云计算对用户端的设备要求最低，操作起来方便简单，这是其他存储方式所不具备的。举例来说，我们平时经常需要更新各种应用软件，有时候为了能够使用最新的软件版本或操作系统，我们也不得不花费大量的时间升级电脑硬件。最令人头疼的是，有朋友发来了某种格式的文档，我们也不得不根据这个文档的需求来下载安装某个应用软件。为此，电脑上装了一大堆可能只会用一次的应用软件，大大降低了电脑的运行速率。而且，为了阻止下载时随时可能侵袭而来的病毒，又必须要装载杀毒软件和防火墙。

这些烦琐的程序加在一起，就像遇到了接踵不断的麻烦。有时候就为了看一个几百字的文档，却需要花费几十倍的时间。而你恰巧又是一个电脑新手，这种体验对你来说绝对是一场噩梦！

这时候，云计算的优势就体现了出来，它将带给你全新的简洁体验。首先，你只需要有一台电脑，它可以上网；其次，你的电脑上有一个你喜欢的浏览器。接下来你只需要在浏览器中键入URL，然后就可以尽情享受云计算带给你的无限乐趣了。

你也可以在浏览器中直接编辑存储在“云”的另一端的文档，随时可以与朋友分享信息，而你再也不用担心电脑里的软件是否是最新版本，再也不用为恼人的病毒而发愁。在“云”的另一端，有专业的IT人员帮你维护硬件，帮你安装和升级软件，帮你防范病毒和各类网络攻击，帮你做你以前在个人电脑上所做的一切。这就如同你请了一个忠实能干的管家，你只需要在家里安静地享受舒适与安全，他会替你打理一切琐事。

轻松共享数据

云计算可以轻松实现不同设备间的数据与应用共享。这一点，就大大地减少了数据间转移的麻烦。举个例子，我们的手机里存储了几百个联系人的信息，可是当买了新手机之后，不得不把旧手机上的号码转移到新手机上，这时候需要同步。家里的电脑和办公室的电脑也是如此，需要经常地进行同步才能保证信息不遗漏。

由于不同的设备其数据同步的方法繁多，操作起来也比较复杂，想要在这众多不同的设备之间实现“最新联系清单”的愿望，就要付出大把的时间成本。而云计算能够让一切都简单起来。

在云计算的网络应用模式中，数据只有一份，保存在“云”的另一端，你的所有电子设备只需要连接互联网，就可以同时访问和使用同一份数据。换句话说，这份数据永远是最新的。

可能无限多

云计算为我们使用网络提供了无限的可能性，这种可能性几乎是你任何的想象都可以达到的。比如，你打算在天气晴朗的时候和家人驾车出游，这时候你想要查看一下自己所在位置的交通，你只需要用手机连入网络，就可以快速清晰地看到你目前所在地的卫星地图，并从上面及时便捷地掌握交通状况，你还可以在出行的途中快速地查询自己预设的行车路线，还可以和网络上的好友实时交流，找到他人推荐的附近最好的餐馆、酒店和美丽的风景，你也可以活动下手指，预订目的地的酒店……这一切都是如此美妙，当然了，我们在享受体验的时候不要忘记分享，你还可以把自己刚刚拍摄的照片或视频进行剪辑，分享给正在远方关注你的亲朋好友。

如果没有云计算，单单只是使用个人电脑或手机上的客户端应用，这些无限可能的享受就体验不到了。不但如此，就存储量来讲，个人电脑或其他电子设备就显得又笨拙又“小气”，因为硬盘的存储空间总是有限的，也无法为我们提供无限量的计算能力，但“云”可以轻松实现这一点。在“云”的另一端，有成千上万台甚至更多的服务器组成的庞大集群在支撑着，我们几乎无法想象“云”的另一端到底能承载多少数据存储和计算，因为那几乎是无限的。

云计算很好地体现了互联网的精神实质——自由、平等和分享。在大数据时代，云计算已经展现出了无穷的生命力，在未来，也将更多地改变和影响我们的工作和生活，大大地提高我们的生活质量。

无论你是一名普通网络用户，还是企业的员工，又或者你是一名IT管理者，一个刚刚接触电脑的新手，只要你使用了云计算，都能亲身体验到这种显而易见的改变。

☆中国的云计算产业链

任何一个产业都存在一个产业链，分为上游、中游和下游环节，云计算产业也是如此。中国的云计算产业生态链目前正在政府的监管下快速地构建，其规模每年都在成倍增加。

云计算产业链的构成主要包括四个部分：

第一部分：云计算服务提供商。

第二部分：软硬件、网络基础设施服务商。

第三部分：云计算咨询、规划、交付、运维、集成服务商。

第四部分：终端设备厂商。

这四大部分共同构成了云计算的产业生态链，为政府、企业和个人用户提供实用和细致的服务。

☆云计算产业链之所以备受关注，主要基于五大方面

云计算扩展的投资价值

与传统模式相比，云计算简化了软件、业务流程和访问服务的环节，这方面的改进有力地帮助企业优化了他们的投资规模，并使操作更加便利。

混合云计算的出现

云服务是一个新开发的业务功能，企业使用云计算（包括私人和公共）来补充他们的内部基础设施和应用程序，这些服务可以简化一些业务流程，优化企业的业务性能。

以云为中心的设计

目前，越来越多的企业采用云技术，他们将组织设计作为云计算迁移的元素。这是一个很大的趋势，伴随着云计算的扩展，这种趋势会逐渐扩展到不同的行业。

移动云服务

云服务的未来一定是移动，而数量上升最显著的平板电脑、iPhone和智能手机等移动设备，则在移动中发挥了更多的作用。越来越多这样的设备被用来开发企业的业务流程、通信等功能。

云安全

任何数据都存在安全问题，人们当然也十分担心存在云端的数据是

否安全。用户期待看到更安全的应用程序和技术，而正是基于这个原因，未来将会有许多新的加密技术和安全协议出现。

大数据的出现为IT行业的发展提供了新的契机，就像云计算。不过，同时也伴随着更多的挑战，而且已经迎来一个激烈竞争的环境，亟待政府进一步出台规范措施。

☆1+1＞2

从18世纪以蒸汽机为先导的第一次工业革命，到19世纪末20世纪初以电器技术为先导的产业革命，再到从50年代到80年代以电子和信息技术为先导的信息技术革命，人类经历了三次经济的巨大变革。

时至今日，大数据正引领信息革命的一个新时代。

早就有专家预测，在未来的几年里，伴随着互联网、物联网的发展以及数据信息的不断壮大，可能会有2100亿个RFID或者集群出现。我们可以试想一下，在日常身处的环境中，假如未来的移动互联、物联网变成现实，那我们的生活将会是另外一番场景。到那时候，包围在我们身边的将会是传感器、数据采集装置和更加庞大的数据量。

问题是，如果这些数据量不被分析处理，它们就仅仅是数据，没有意义，没有价值。想要实现数据的价值，就要把它们变成信息、智能和商业的价值。而这才是大数据的真正意义所在。

在三年前，大数据这个词在中国几乎还没有人提起过，而现在，市场上到处充斥着大数据，人们谈论的也都是大数据。大数据的发展是清晰可见的。发展到现在，除了数据本身发生了改变之外，云计算的出现也使数据变得更加分散了。这种趋势是对传统数据库的一大挑战，因为传统数据库已经难以满足市场对于海量数据的需求，而且随着数据的多样化、对数据的需求加快，传统数据库更是显得举步维艰，这时候，各种各样的解决方案纷纷现身。

由于大数据本身就是一个问题集，在众多的解决方案中，最重要和最有效的技术是云计算，两者结合起来，将产生1+1＞2的效果。而这也是人们公认的处理大数据集最有效的分布式处理手段。云计算为大数据的处理提供了基础架构平台，大数据应用可以在这个平台上运行，双方密不可分，互相保障。

对于大数据给云计算带来的影响，美国一位IT公司的技术总监贝斯特表示，大数据对云计算的影响只表现在私有的云架构上，对于公有的云架构，对数据仓库没有影响。因为企业的CIO不会无缘无故把财务数据或者客户数据放到云上，因为那是一件极度危险的事情。而私有的云架构则不同，它对于数据仓库的影响有两点：

第一，通过私有云，可以巩固数据集，减少利用率不足的问题；

第二，可以通过灵敏的方式将数据集成，实现业务价值。

这保证了双方不会发生任何冲突，反而起到了互相补充的加强作用。

☆云计算与大数据的区别——应用的分工

概念的不同

从宏观的概念上来讲，云计算改变了IT，而大数据则改变了业务。同时，大数据必须有云作为它的基础架构，才能得以顺畅推广并体现出强大的实用价值。

目标受众的区别

双方的目标受众也是不一样的，云计算代表着一种IT层面的解决方案，是面向CIO的；而大数据则是一种战略构架，是面向管理者和业务层的，它能让我们在业务上展示出更强大的竞争力，完全提升综合实力。

在中国的发展

☆云计算在世界的发展状况

说起云计算的发展前景，各大互联网巨头就像看到一支飞速上升的A股一样，对于云计算充满了信心。为此，很多公司开始调整未来发展战略。

例如，亚马逊使用弹性计算云（EC2）和简单存储服务（S3）来为企业提供计算和存储服务。其中，收费的服务项目包括存储服务器、带宽、CPU资源和月租费。月租费的含义与电话月租费类似，存储服务器、带宽按容量收费，CPU则根据时长(小时)运算量收费。亚马逊把云计算做成了一个类似于移动通信的生意，而且只花了不到两年的时间。

根据某第三方机构提供的数据，在亚马逊上注册开发的人员达到了44万人，其中有很多是企业级用户。而亚马逊与云计算相关的业务收入额也已经达到了1亿美元。在亚马逊所有增长最快的业务中，云计算就是其中之一。

要说云计算使用者最多的网站，当数谷歌。这一点是毋庸置疑的，因为支撑谷歌搜索引擎的，是分布于200多个地点、超过100万台服务器的基础设施，而这些设施的数量正在迅猛增长。无论是谷歌地球、地图、Gmail还是Docs等等，同样都使用了这些基础设施。从这一点来说，谷歌公司确实是非常厉害的，因为采用了Google Docs之类的应用，用户的数据都会保存在互联网上的某个位置。而且，用户可以通过任何一个与互联网相连的系统便捷访问这些数据。

谷歌还有一点值得称颂的就是它的“分享”精神。目前，它已经允许第三方在谷歌的云计算中通过Google App Engine运行大型并行应用程序。而且，早先它就已经以学术论文的形式对外公开发表其云计算的三大法宝：GFS、MapReduce和BigTable。在美国、中国等一些高校，谷歌也开设了关于如何进行云计算编程的课程。

我们再来看IBM。2007年11月，IBM推出了改变游戏规则的“蓝云”计算平台，这个云计算平台为客户带来了即买即用的体验。它包括一系列的自动化、自我管理和自我修复的虚拟化云计算软件，使来自全球的应用可以访问分布式的大型服务器池，使得数据中心在类似于互联网的环境下运行计算。

作为世界PC软件先导的微软公司同样紧跟云计算的步伐，在2008年10月的时候正式推出了Windows Azure操作系统。Azure(中文译为“蓝天”)是继Windows取代DOS之后，微软的又一次颠覆性转型。通过在互联网架构上打造新的云计算平台，让Windows真正由PC延伸到“蓝

天”上。微软拥有全世界不计其数的Windows用户桌面和浏览器，现在它将它们连接到“蓝天”上。为Azure的底层提供支撑的是微软全球基础服务系统，由遍布全球的第四代数据中心构成。

☆云计算在中国的发展实况

关于互联网的未来发展方向，中移动前董事长兼CEO王建宙就认为，毫无疑问的，必定是云计算和互联网的移动化。也就是说，中国互联网事业的发展前途，在很大程度上就取决于云计算在中国的前途。

云计算在中国的发展历程：

2008年5月10日，IBM在中国无锡太湖新城科教产业园建立的中国第一个云计算中心投入运营；

2008年6月24日，IBM在北京IBM中国创新中心成立了第二家中国云计算中心——IBM大中华区云计算中心；

2008年11月28日，广东电子工业研究院与东莞松山湖科技产业园管委会签约，广东电子工业研究院将在东莞松山湖投资2亿元建立云计算平台；

2008年12月30日，阿里巴巴集团旗下子公司阿里软件与江苏省南京市政府正式签订了2009年战略合作框架协议，在南京建立国内首个“电子商务云计算中心”，首期投资额达上亿元人民币；

2009年，世纪互联推出了CloudEx产品线，包括了完整的互联网主机服务CloudEx Computing Service,基于在线存储虚拟化的CloudEx Storage Service，供个人及企业进行互联网云端备份的数据保全服务等等系列互联网云计算服务。

从2010年开始，中国的云计算更是被纳入了国家重点工程，获得了政策、资金和技术上的倾斜支持。比如作为中国对云计算探索研究较早的中移动研究院，目前已经成功完成了云计算中心的试验。

对于“云安全”，中国企业创造的概念在国际云计算领域算是独树一帜。我们的“云安全”的概念是：云安全通过网状的大量客户端对网络中软件行为的异常监测，获取互联网中木马、恶意程序的最新信息，推送到服务端进行自动分析和处理，再把病毒和木马的解决方案分发到每一个客户端。

中国对于“云安全”的策略构想是：使用者越多，每个使用者就越安全，因为如此庞大的用户群足以覆盖互联网的每个角落，只要某个网站被挂马或某个新木马病毒出现，就会立刻被截获。

“云安全”的发展就像一阵龙卷风，很快席卷了各大安全应用企业。瑞星、趋势、卡巴斯基、MCAFEE、SYMANTEC、江民科技、PANDA、金山、360安全卫士、卡卡上网安全助手等都相继推出了云安

全解决方案。例如瑞星，基于云安全策略开发的2009新品每天拦截木马攻击的数量就达到了几百万次，而其中仅在2009年1月8日这一天就拦截了765万余次。

据悉，云安全可以支持平均每天55亿条点击量查询，从这些点击量中每天收集2.5亿个样本加以分析，根据这个庞大的资料库，第一次命中率就可以达到99%。借助云安全，趋势科技现在每天阻断的病毒感染数最高可达1000万次。

其实云安全的核心思想并不是第一次被提出，早在2003年，中国的云计算专家刘鹏就曾提出过反垃圾邮件网格，这与云安全的思想非常相似。刘鹏当时的想法是，针对网络垃圾邮件的泛滥，仅靠技术手段无法很好地自动过滤，因为邮件过滤所依赖的人工智能方法并不成熟。那么，根据垃圾邮件“将相同的内容发送给数以百万计的接收者”的特征，就可以建立起一个分布式的统计和学习平台，以大规模用户的协同计算来过滤垃圾邮件。

这个方法是怎么实现的呢？

首先，用户需要在电脑上安装一个客户端，然后就可以为收到的每一封邮件计算出一个唯一的识别码，就像“指纹”，通过比对“指纹”就可以统计出相似邮件的副本数，当副本数达到了一定的数量，就可以判定哪些邮件是垃圾邮件。

其次，由于互联网上多台计算机比一台计算机掌握的信息更多，因而可以采用分布式贝叶斯学习算法，在成百上千的客户端机器上实现协同学习过程，收集、分析并共享最新的信息。

由此看来，用大规模统计方法来过滤垃圾邮件的做法确实要成熟很多，而且误判率低，具有很强的实用性。从思想核心来看，反垃圾邮件网格也更真实地体现了网格思想，因为每个加入系统的用户在作为服务对象的同时，也是完成分布式统计功能的一个信息节点。随着系统规模的不断扩大，系统过滤垃圾邮件的准确性也随之提高。

这既是一个服务的过程，也是反哺并提升技术进步的过程。反垃圾邮件网格就像一张“天网”，充分利用了分布于互联网中的千百万台主机协同工作，由此构建起一道拦截垃圾邮件的天然屏障。

IEEE Cluster 2003国际会议曾把反垃圾邮件网格选为杰出网格项目，并且在香港作了现场演示，引起了世界各地广泛的关注。在2004年网格计算国际研讨会上，还作了关于反垃圾邮件网格的专题报告和现场演示，很多邮件服务商表现出极大的兴趣。而中国最大的邮件服务提供商网易创办人丁磊，对此更是非常重视。

所以我们再回到之前的看法，垃圾邮件尚可如此处理，那么病毒、

木马等也是同样的道理，这样看来，与云安全的思想就很接近了。

对于大数据在中国的兴盛，中国有一位资深的大数据研究者表示出了这样的看法：“现在中国所谓的大数据公司，都还是在以互联网思维理解大数据，而非真正的大数据思维，未来还有着大量的创新空间。”

这就是说，大数据在中国的发展，虽然已经有了一定的进展，但还远远不是我们希望看到的。就像人们对于第一次信息技术革命的预测一样，谁也没有想到，现在会是信息技术的天下。很显然，中国的投资人们可不想再等上十年八年，到那时候，他们俨然已经错过在这场大变革中最早的布局机会，市场早被他人垄断，到时候黄花菜都凉了。

对于目前市场上的大数据公司，我们大体可以将其分为三类：

第一类，拥有大量的用户信息，通过对用户信息的大数据分析解决自己公司的精准营销和个性化广告推介等问题。如亚马逊、谷歌和Facebook。

第二类，通过整合大数据的信息和应用，给其他公司提供“硬件+软件+数据”的整体解决方案。如IBM和惠普。

第三类，通过出售数据和服务更有针对性地提供单个解决方案。这一类基本上是新兴的创业公司。

需要特别指出的是，作为第三类的新兴创业公司，它们将大数据进行商品化，这会引发继门户网站、搜索引擎、社交网络之后的新一波创业浪潮和产业革命，并且一定会对传统的咨询公司产生强烈的冲击。

不过，如果我们仅仅把大数据的影响力框定在对传统咨询公司的冲击之上，似乎也有些小看了它的威力。大数据分析与传统的数据分析、数据挖掘具有一定的延续性，关键不同在于其分析的数据量更为巨大，且多为非结构化数据。譬如很多段小视频，或是电子商务里的各种评价、晒单等等。这与传统数据分析多利用cookie获取诸如用户每月登录某网站几次等结构化的数据，在技术处理方式上有着很大的不同。

但是，从大数据分析在根本上要做的事情来说，它仍然是要在这些大量的数据中进行分析，得出一些对商业决策有帮助的pattern(模式、方法)。它的应用空间会非常广泛。

对于现在国内一些广告平台公司、市场公司都纷纷上马大数据业务的现象，我们不得不指出的是，它们很多并不真正了解自己的需求，也不明白大数据意味着什么。在实际应用中，大数据主要包括了大交易数据、大交互数据和大机器数据三类。第一类大交易数据已存在多年，从传统银行、电信的交易数据到各类网银支付数据都包括在内；第二类大交互数据，则主要是指来自脸书、推特、微博等社交网络的非结构化数据；第三类大机器数据，则是指由物联网内各种传感器所产生的数据。

如果不是真正做这三类工作并且处理它们之间关系的，即便上马了大数据，也只能是形似神不似，花钱不少，但获得不了多少实际的价值。

现在，中国的老板们眼中都看到了“数据财富”的可贵，也大都在采取一些行动。这是一个庞大的朝阳产业，仅仅在大数据自身的产业链上，就可以分为数据采集、数据清洗、数据分析和垂直行业算法四个环节。但由于中国的市场规模才刚刚起步，分工还没有细化，中国的大数据先行者必须从头开始，甚至在结合国外经验的基础上，要摸着石头过河，才能一步步总结出符合中国国情的大数据应用战略。

阿里巴巴：云帝国构想

我至今仍然对马云说过的一句话印象深刻："再不动就要死！"阿里巴巴从创立开始，就始终遵循一种"不动即死"的战略原则。1999年，马云创立了作为企业对企业的网上交易平台阿里巴巴。2003年，又投资1亿元人民币建立了淘宝网。2004年，阿里巴巴开始推出支付宝服务，面向中国的电子商务市场提供基于中介的安全交易平台。

淘宝和支付宝，已成为阿里巴巴在电商领域的两大互补性支柱，一跃成为全中国最强大的电商企业。但是马云没有停止扩张的步伐，阿里巴巴仍在继续"动"。他先是购入高德地图，投资新浪微博，而且还增资UC。在十周年活动后，马云宣布卸任，但阿里巴巴的"云帝国"才刚刚开始。

☆传播渠道——天下网商

有了传播渠道，品牌的扩散就有了保证。传播也是信息辐射的重要平台，因此越是信息丰富的社会，媒体的重要性也就越突出。阿里巴巴在2010年和浙江出版联合集团倾力打造了一家新媒体《天下网商》，专门为其电子商务领域提供信息传播服务，这成为了阿里巴巴品牌战略的一大标志，也对它的品牌地位有了范围更广的提升。

☆核心数据源——旗下的拳头产品

好的产品才是成功的基础，阿里巴巴旗下的所有产品几乎都是一个强大的数据源。从1999年以来，包括阿里巴巴黄页、淘宝网、天猫、一淘、聚划算、阿里旺旺等产品相继崛起，独树一帜，占据了行业领先的地位，为阿里的品牌战略提供了无数充实的内容，也为阿里的大数据战略提供了坚实的核心数据，成就了马云的全网络战略梦想。

☆核心技术——阿里云

技术永远都是生产力的发动机。拥有了过硬的技术，才能收集海量数据并发现商机，再提供全方位的产品。在这方面，阿里巴巴不惜重金，一直在不遗余力地提升自己的云计算能力。2009年，阿里云计算成立；次年4月份，phpwind正式进入阿里云计算有限公司，成为阿里云计算的战略性重点产品；当年8月25日，阿里巴巴宣布和专门服务于eBay商家的第三方工具开发商Auctiva达成收购协议，进一步提升了阿里云的辐射能力。

☆数据资产——生活地图

数据就是财富，但它的收益工具是什么？阿里巴巴以一种"生活地图"的实用理念为用户提供数据服务。2010年8月，阿里巴巴向易图通注资3500万美金，成为该公司最大的股东；2012年10月份，阿里巴巴推出

了淘宝地图服务；2013年5月，又以2.94亿美元购入了高德地图28%的股份。

☆数据流——流量统计技术提供商

在2009年，阿里巴巴就投资了万网并和其结盟，投资金额是5.4亿元；2011年，阿里巴巴又完成了对流量统计技术提供商CNZZ的收购；但这并没有结束，仅过了两年，也就是2013年的11月，阿里巴巴宣布收购移动互联网统计分析公司友盟。要知道，友盟公司掌握了超过10万移动应用的数据，这些数据被认为对分析用户行为和移动电商广告精准营销具有重要的价值，属于价值最高的数据流。

☆数据收集——搜索引擎

在搜索引擎方面，国外有谷歌，国内有百度。2010年10月，阿里巴巴也开始介入这一领域，作为战略投资者与云峰基金联手投资搜狗，共计投资1500万美元。虽然仅过了一年多的时间，搜狐公司就以2580万美元回购了此前阿里巴巴所持有的搜狗10.88%的股份，但早在投资搜狗的同时，阿里巴巴自己的一淘购物搜索也成立了，并且功能更专一、价值链更集中。

☆数据扩张——阿里巴巴的国际化

阿里巴巴在奠定国内第一电商的地位后，仍然不断扩张。2010年6月25日，阿里巴巴收购了美国电子商务SaaS提供商Vendio Services，这是阿里巴巴第一次在美国市场上进行收购活动。同年11月16日，阿里巴巴又宣布收购深圳市的一达通企业服务有限公司，使自身的阿里系产品无处不在。

☆移动互联——数据改变生活

在移动互联网领域，没人想落后于人，这是数据真正改变生活的战场，控制了移动互联平台，就意味着控制了数据来源，拥有了对全数据的全方位的收集能力。从2012年11月到2013年4月底，在不到7个月的时间内，阿里巴巴战略投资陌陌，收购虾米网和墨迹天气，以5.86亿美元购入了新浪微博18%的股份，在移动互联领域做出了一系列的战略举措，为阿里云帝国的数据来源造血。

综上所列，阿里巴巴的云帝国布局始终围绕着三个方面运行：品牌战略、大数据战略、帝国战略，三者互相配合，互为基础，而且层层递进。到目前为止，马云已经把阿里巴巴的大数据战略布置完成，这个在十几年前诞生于杭州的中国小公司，伴随着电子商务在中国的发展，一路走来，今天已经成为一家拥有包含淘宝、阿里云、天猫、聚划算、小企业业务等支柱业务的庞大帝国。

阿里巴巴的现实与梦想，也预示了中国的大数据产业未来的发展方

向。在2012年的一次调整中，阿里巴巴正式组建了7大事业群，并且以大数据为基础，明确提出了CBBS市场体系，即由淘宝的C，至零售商、渠道商B，再到生产制造商B，而S是一种服务，面向整个产业链上的B和C服务。

马云在写给员工的邮件中说：“我们设立CBBS大市场体系，是为了在今天和未来的严峻经济形势下，完善自我，全面提升集团对小企业和消费者的服务能力，帮助小企业渡过生存和成长难关，同时让更多的消费者受益于互联网时代的丰富生活。最终促进一个开放、协同、繁荣的电子商务生态系统。”

根据这个运作体系，我们能够清晰地看到大数据思维是如何在电子商务的产业链中成功发挥作用的。例如在服装定制行业，有需求的消费者只需要在一家淘宝商户那里填写身高等几个关于体型的问题，并且对图画中的身形进行选择，IT系统就会基于存储的会员数据，自动地生成匹配准确率达90%的数据，然后商家将数据进行分析，通过IT系统把每个部分的尺寸、用料信息发给阿里巴巴平台上的供应商；供应商在接到订单信息以后，采用相应的原料就可以立刻生产。

在这个体系下，阿里巴巴建立了一个完整的由数据驱动的电子商务生态链，从消费者和渠道商在个人消费平台（比如淘宝或天猫）的网上交易开始，由企业贸易平台（阿里巴巴国际业务和阿里巴巴小企业业务）不断地供货。最后，渠道商把消费者的个性需求反馈到制造商，为消费者按需生产、个性化生产。

在这个过程中，全部的交易模式都离不开“阿里云”的保障。阿里巴巴的电子商务服务商将同时为这个产业链条上的每一个环节提供周到及时和高质量的服务。坐拥“阿里云”的魔法之剑，阿里巴巴的云帝国已经成形。

腾讯：大社交战略

腾讯公司从成立起，就始终是移动互联和社交领域的数据大鳄，有多少人在用QQ和微信，就有多少人免费甚至花钱成为腾讯的数据源。在社交工具方面，腾讯的优势无可匹敌，早在10年前就奠定了它在国内遥遥领先的“社交帝国”的领头羊位置。

在2012年的一次经济峰会上，腾讯董事会主席兼CEO马化腾说：“社交媒体的广告现在还是一个没有挖掘的宝藏，这是非常值得我们思考的。”这说明腾讯公司在大数据时代对于潜在市场的预见力和控制一切的野心已足够成功，却仍需努力。

思考的结果是，仅仅在两个月以后，腾讯便宣布架构重组，把现有的业务重新划分成了企业发展事业群(CDG)、互动娱乐事业群(IEG)、移动互联网事业群(MIG)、网络媒体事业群(OMG)、社交网络事业群(SNG)和技术工程事业群(TEG)，并且成立了腾讯电商控股公司(ECC)专注运营电子商务业务。

在公司内部的邮件中，马化腾对于“社交领域”做出了一个新的定义：“我们要强化大社交网络。”具体做法就是，将即时通讯平台QQ与两大社区平台QQ空间、朋友网整合成为一个大的社交网络事业群，然后形成更具有规模效应的社交网络平台，这就是腾讯的大社交战略，推行保障就是它已经具备的强大的大数据能力。

☆挖掘社交的宝藏

对于腾讯的这次重组，一位高级分析师评价说：“腾讯公司盯准的是在当下及未来的中国市场最有前途的商业模式，它最拥有胜算的市场是网游及SNS。也就是说，腾讯公司准备在大数据领域挖掘社交的数据宝藏，并追求建立它的垄断地位。”

在这次大调整之后，腾讯的社交网络事业群空前强大，囊括了即时通讯部门（含QQ、企业QQ等腾讯核心产品）、QQ空间和朋友网。而且，据它自己公布的数据显示，腾讯IM的活跃账户数已经超过了7亿；QQ空间的活跃账户数也高达5.5亿；朋友网的活跃账户数则超过了2亿。

这是一个什么概念？说明它基于QQ空间和朋友网的广告系统的日流量可以高达几十亿。这么优质的资源，如果空置在互联网平台上，岂不是白白浪费掉了？对于腾讯而言，充分地开发和挖掘这一块黄金宝地，是下一步的重点。

在宣布重组之前，腾讯公司就先公布了它的社交化营销平台，这说明马化腾早已为这场战争的胜算做好了充足的准备，也揭开了国内的大数据公司转向广告层面来实现数据资产变现的大幕。与此同时，这一场

变革还有另一层意义：伴随着网络媒体、广告产品及腾讯营销方法论的全面升级，腾讯公司希望在未来创造和控制一个不亚于搜索引擎的市场。

对此马化腾说：“我们已做了很多的努力，包括传统门户网站，社交媒体的广告是一个还没有挖开的金矿，像国外的脸书通过用户在社交中的行为把人进行划分、针对人的不同属性做广告，值得我们国内的公司去学习。”

☆脸书（Facebook）：最好的老师

脸书是全世界最大的社交网络，它的市值高达869亿美元，其中超过85%的收入都来自于广告。广告的价值体现就在于这个平台的数据流量（点击量和人气）；与之相比，腾讯是中国最大的社交网络，市值约560亿美元，但它90%的收入都来自个人用户的增值服务，而不是广告。

所以，腾讯公司把学习的榜样对准了脸书，重点挖掘社交网络的广告价值。现在，虽然国内很多巨头都在学习脸书，但并没有学到“核心智慧”。比如开心网和人人网（以前的校内网），它们的主要收入和脸书一样都来自广告，但两者的区别是非常明显的，具有本质的不同。中国的社交网站营收过于依赖投放式的广告，收入也不理想，显然脸书不是这样做的。

脸书并不只提供传统广告（粗狂的投放），它鼓励各大品牌公司与用户的沟通，向用户们讲述自己的品牌故事。脸书的CEO扎克伯格相信，这样的互动方式比其他的任何网络广告都更具有亲和力。于是，脸书根据用户的注册资料信息去推送相对精准的广告内容，并且使用社交网络构建的人与人之间的关系也有利于传播广告信息，把广告转化为有效的内容，来为各种广告商提供高效率的广告解决方案。它的精准广告技术绝不会为了收入去牺牲自己的产品，反而赢得了企业的喜爱。

☆腾讯的大数据营销

针对脸书的经验，腾讯制定了适合中国市场的“大数据营销”。对于什么是大数据，腾讯公司的解释别有新意，它把大数据定义为：如果信息的复杂性、大小已经大到我们很难用一种普通数据工具去描述的时候，那么不管在收集、管理还是预算领域，都可以把它称之为大数据。

腾讯公司拥有全中国最丰富和最庞杂的数据量，它要做的就是把自己拥有的社交网络所积累下来的数据，经过分析和挖掘，进行精准归类：“我们的用户都是一些什么样的人？”然后确定营销的方向。

在这个营销模式的平台用户管理界面上，企业能够清晰地了解自己的用户群体特征，通过分析海量数据，对广告进行精确匹配，这正是脸

书的模式。虽然效果如何还不得而知，但这注定是一项长期的工作。中国的社交网络拥有巨大的潜力，你只知道一件事就行了：大数据，腾讯早就在路上。

360：最大数据中心

在信息安全领域，似乎没有哪家公司比360更引人注目。在利用大数据分析重新构建企业的信息安全框架方面，奇虎360走在了前面，而且越走越快。

比如，在2011年，360就与中国电信量子数据中心达成了全面合作意向，陆续在该中心投放2000台容纳大量服务器的标准机柜，作为其在全国范围内最大的单体数据中心。它的机房面积超过了1万平方米。该中心的主要职责就是为互联网内容应用商提供服务器托管、日常维护和维修等服务，保证其服务器的正常运转。通过这些服务器，360得以面向全球的互联网数据。

360说："我们现在使用的运营商资源非常大，我们企业发展靠的就是带宽、机房和服务器，这个数据中心的机房非常完备，堪称五星级，我们将把重点业务放在这里，包括我们的全球'一对一'客服。"

在对数据的收集和信息安全方面，360是国内的佼佼者。到2013年的8月份，它的总数据量（安全卫士）已经达到了16.85T，每天的数据量达81.24GB，每天总请求数超过31亿次，每天拦截漏洞攻击30万次，每天拦截IP攻击50万次。

360是如何通过如此庞大的数据来判断潜在的安全风险的？

它的网站卫士实时数据分析平台主要是由Scribe、Storm、Inotify、Rsync这四个模块（技术分区）构成，离线数据的分析平台则由Scribe、Hadoop和M/R构成。360组合式地来使用这些技术工具，每天拦截互联网上的各种木马通信。公司的技术人员做过一个统计，他们发现52%的攻击来自可疑通信，14%为外挂插件，10%则来自后门通信，19%来自其他攻击方式的攻击。

360的相关部门负责人表示，企业和个人网站面临的威胁可以通过外包建站、开源程序、第三方接入、安全意识、同行竞争五个方面进行评估。这是经验之谈，也是对目前的网络信息安全的全方位总结。

在运营战略上，360与其他巨头相比，走出了一条富有特色的道路。比如，它的杀毒是完全免费的（国外的卡巴斯基等杀毒软件则收费昂贵），通过提供高品质的免费安全服务，来为用户解决上网时遇到的各种安全问题。

360是免费安全的首倡者，它对此的战略认识是：互联网安全应该像搜索、电子邮箱和即时通讯一样，成为一项免费的基础服务。

这让它赢得了无数"金字塔"中下层的用户，拥有庞大的用户数据量。同时，它还开发了具有全球规模和领先技术的云安全体系，来帮助

自己的安全卫士和杀毒软件为用户更好地服务，保护用户的上网安全，当然也就保护了自己挖掘出来的这些数据量。

作为中国最大的互联网安全公司之一，360拥有高水平的技术团队，旗下360安全卫士、360杀毒、360安全浏览器、360安全桌面、360手机卫士等系列产品集成了一个产品群，几乎涉及互联网用户的所有需求，也使它自己成为了国内当仁不让的网络安全领先品牌。

根据2013年的数据统计，360的个性化起始页和其子页面的日均独立访问用户接近2亿人，日均点击量超过了6亿；PC端产品和服务的月活跃数达到4.61亿，旗下产品的用户渗透率达到96%；使用360手机卫士的用户总数达到了约3.38亿，而其市场渗透率更是高达70%。另外，使用360浏览器的月度活跃用户达到了3.3亿，用户渗透率也接近了70%，在国产浏览器行业处于遥遥领先的位置。最后，360的搜索引擎还具有完全自主的知识产权，拥有18%以上的稳定的市场份额，成为了中国搜索市场的重要参与者。

从2006年成立以来，360只用了五年时间，就重新定义了互联网安全，改写了市场格局。

- 模式：360是免费安全的首倡者，主张安全作为互联网的基础服务，应该彻底免费，建立了用户完全免费的商业模式；
- 技术：360将自己的搜索引擎技术应用于安全领域，建成了规模、用户数、使用量均在国内处于领先位置的360云安全系统；
- 安全：这是最重要的一点，360用全新的用户体验重新定义了互联网安全，建立了大安全观概念，即安全不仅仅等于杀灭病毒，它还意味着数据安全、隐私安全、账号安全、下载安全，以及电脑无死角的“全健康”。

创新和开放始终是360的DNA，通过用户体验创新、技术创新和商业模式创新，360公司在大数据时代积聚了超过4亿的海量数据源，通过资源共享，既让用户获得优质的安全体验，也让自己的商业伙伴得到了丰厚的回报。

结合我们上面的分析，再来看一下360的产品和服务，你就会有新的体会和认识。

安全卫士：查杀木马与电脑“全保护”

360安全卫士是当前功能更强、效果更好、更受用户欢迎的上网必备安全软件。由于使用方便，用户口碑好，目前，首选安装360的用户已超过4亿。它拥有查杀木马、清理插件、修复漏洞、电脑体检等多种功能，并独创了“木马防火墙”功能，依靠抢先侦测和云端鉴别，可全面、智能地拦截各类木马，保护用户的账号、隐私等重要信息。

目前，木马威胁之大已远超病毒，360安全卫士运用云安全技术，在拦截和查杀木马的效果、速度以及专业性上表现出色，能有效防止个人数据和隐私被木马窃取，被誉为“防范木马的第一选择”。360安全卫士自身非常轻巧，同时还具备开机加速、垃圾清理等多种系统优化功能，可大大加快电脑运行速度，内含的360软件管家还可帮助用户轻松下载、升级和强力卸载各种应用软件。

杀毒：永久免费的强大杀毒软件

360杀毒是中国用户量最大的杀毒软件之一，360杀毒是完全免费的杀毒软件，它创新性地整合了五大领先防杀引擎，包括国际知名的BitDefender病毒查杀引擎、小红伞病毒查杀引擎、360云查杀引擎、360主动防御引擎、360QVM人工智能引擎。五个引擎智能调度，提供全时全面的病毒防护，不但查杀能力出色，而且能第一时间防御新出现的病毒木马。

并且，360杀毒完全免费，无须激活码，轻巧快速不卡机，误杀率远远低于其他杀毒软件。360杀毒独有的技术体系对系统资源占用极少，对系统运行速度的影响微乎其微。360杀毒还具备“免打扰模式”，在用户玩游戏或打开全屏程序时自动进入“免打扰模式”，拥有更流畅的游戏乐趣。360杀毒和360安全卫士配合使用，是安全上网的黄金组合。

搜索：安全、干净和有效竞争

360搜索是具有自主知识产权的搜索引擎，包含网页、新闻、影视等搜索产品，为用户带来更安全、更真实的搜索服务体验。360不仅掌握通用搜索技术，而且独创PeopleRank算法、拇指计划等创新技术。目前已建立由数百名工程师组成的核心搜索技术团队，拥有上万台服务器，庞大的蜘蛛爬虫系统每日抓取网页数量高达十亿，引擎索引的优质网页数量超过数百亿，网页搜索速度和质量都已经达到先进水平。

安全桌面：桌面安全和便捷应用

360安全桌面目前汇集了数十万款酷炫互联网应用，分为游戏、视频、小说、音乐、购物、生活、时尚娱乐、实用工具、投资理财、图片、社交、新闻等十余种类别，几乎覆盖了所有主流热门应用。它所有的应用都秉承“绿色”原则，无毒无插件，均经过360安全中心的认证，用户在点击下载后即可一键直达想要的应用。

安全浏览器：让用户放心上网

360安全浏览器是互联网上安全好用的新一代浏览器，拥有国内领先的恶意网址库，采用云查杀引擎，可自动拦截挂马、欺诈、网银仿冒等恶意网址。独创的“隔离模式”，让用户在访问木马网站时也不会感染。无痕浏览，能够更大限度保护用户的上网隐私。

重要的是，360安全浏览器体积小巧、速度快、极少崩溃，并拥有翻译、截图、鼠标手势、广告过滤等几十种实用功能，已成为广大网民上网的优先选择。

极速浏览器：提供极速的浏览体验

360极速浏览器是一款极速、安全的无缝双核浏览器。它基于Chromium开源项目，具有闪电般的浏览速度、完备的安全特性及海量丰富的实用工具扩展。此外，为了更适合国内用户使用，它加入了鼠标手势、超级拖拽、恢复关闭的标签、地址栏下拉列表等实用功能，配合原Chromium的顺滑操作体验，让您浏览网页时顺畅安心。

手机卫士：隐私保护与骚扰拦截

360手机卫士是一款完全免费的手机安全软件：能有效拦截垃圾短信和骚扰电话，让手机恢复宁静空间；联网云查杀恶意软件，实时监控软件安装和联网，彻底杜绝恶意扣费侵害；加密重要联系人的通讯记录，防止个人隐私泄露；系统一键清理，轻松为手机运行加速。还有归属地显示和查询，自动加拨IP节约话费，无痕短信等等众多功能，不仅为用户带来全方位的手机安全及隐私保护，也让用户使用手机更加方便快捷。

手机助手：智能手机的优质助手

360手机助手是Android智能手机的资源获取平台，提供海量的游戏、软件、音乐、小说、视频、图片，通过它轻松下载、安装、管理手机资源。所有提供信息资源，全部经过360安全检测中心的审核认证，绿色无毒，安全无忧。360手机助手帮助用户用更省流量、更快捷、更方便、更安全的方式获取网络资源，为用户的Android手机注入鲜活色彩。

百度：大数据时代的三层布局

从最初的大数据定义之争，到如今挖掘大数据应用价值、协商合作方向，中国的大数据产业已经进入到了一个务实发展阶段。既引发了政府的关注，也在中国重视大数据的千亿公司之间创造了一个良性竞争的环境。这其中，就包括国内搜索引擎业的霸主百度。

2012年，在一次世界营销论坛上，百度副总裁王湛向人们介绍了百度公司对大数据的认识和布局："在大数据时代，信息量庞大，垃圾和金子都蕴含其中，我们要有非常好的炼金术才可以把金子找出来，这是整个大数据时代最重要的认知。"他认为，大数据是继云计算、物联网之后IT产业又一次颠覆性的技术变革，这对于经济发展、企业的决策、组织和业务流程、个人生活方式都将产生不可估量的巨大影响。人们每分每秒都在产生巨量数据，大数据不仅体量巨大，而且类型繁多。

他说："请相信，百度正在开展一场大数据革命，以对接企业的时代需求。我们已经从数据、工具以及应用三个层面，做好了大数据时代的布局，为广告用户更深入地挖掘数据价值，优化营销决策。"

那么，百度的优势在哪儿？

在数据层面，凭借强大的入口优势，在国内市场成功击败世界巨无霸谷歌的百度公司拥有全中国最大的消费者行为数据库，覆盖了高达95%的中国网民，日均响应50亿次搜索请求，搜索市场占比达87%，同时还有百度联盟，60万联盟合作伙伴每天有50亿次的日均行为产生。

另外，百度的另一主打产品贴吧的日均访问量达5.5亿，百度知道累计解决用户问题1.9亿个，其他比如百度图片、视频、MP3、地图、百科、文库等，每天都在以亿级速度储存并分析用户数据。所有的平台数据已经形成了一个数据集合体，将消费者从需求、搜索、购买，到使用和分享的整个真实的历程全程记录下来。

有了这个优势，百度公司就拥有了巨量的、可以充分反映消费者真实需求的数据。在这个基础上，百度公司建成了百度指数、司南、风云榜、数据研究中心和百度统计五大数据体系平台，来帮助企业实时地了解消费者行为、兴趣变化，以及行业发展状况、市场动态和趋势、竞争对手动向等信息，以便它们能够适时、正确地调整企业的营销策略。

☆技术分析

除了以上五大数据平台之外，百度的技术分析也是领先的，拥有消费者画像、品牌探针等分析方法，帮助广告客户洞察消费者背后的故事，比如他们的兴趣点、地域行为差异、媒体接触点、品牌认知和生活形态等各自是什么，由此系统地形成各种最接近于事实的数据统计和营

销指南，为客户品牌整合及市场战略的制定提供宝贵信息。以强大的技术分析能力做保证，客户就能从百度的数据库受益，不用再凭感觉营销，而是可以做到完全基于数据，进行精确投入。

☆联合商业计划

百度积极提升自己与广告客户的关系，对大数据营销共同进行研究。比如在2012年，百度就与宝洁公司一起推出了一项名为“联合商业计划”的合作，包括以消费者画像为主要内容的市场研究，深入地洞察消费者对品牌和产品的认知，帮助其探索有哪些新的途径和手段可以影响消费者。双方还共同对大数据的品牌营销进行了创新，像“感谢妈妈”活动，和宝洁建设了活动官网，用户可以在地图上标注妈妈的位置，传递对母亲的感激和挂念之情。

在这种双方的深入合作中，百度能够整合贴吧、地图、无线客户端、MP3等全媒体平台推广资源，结合客户的一线门店及销售平台，达到非常好的效果，最大化地利用了彼此的数据资源，也让自己积累了在大数据时代的营销实践经验和对于消费者的认知。

☆数据应用的产品推广

离开具体的产品，大数据的应用就成了无本之木。有了产品，就有了务实发展，大数据的应用才能落地，大数据的价值才能向纵深层面扩展，去整合社会数据，让消费者和用户受益，真正成为产业的一个不可分割的重要部分。百度在产品推广方面的成果堪称国内做得最好的，包括语音助手、以图识图、机器翻译、搜索、广告等产品方面，都成功应用了大数据，极大地提升了它的产品价值。

语音识别

百度依靠海量数据，使用深度学习技术，使语音识别准确率大大提升，达到了业内比较高的水平。

搜索

百度本身就是以搜索起家的，因此它在搜索方面的产品一直处于国内领先地位。通过挖掘和整合大数据，百度现在提供了更加智能的搜索服务。比如我们在百度搜索“上海的人口”，会直接反馈给我们一个准确的数字，而不需要再打开别的网页去查找，非常方便。

搜索产品的进步，不但已经创造出了互联网广告价值，还带来了IT之外的无限价值，就是利民和便民的社会应用。比如为民众提供了满足自身的民生需求的医疗、交通搜索服务。我们可以在搜索引擎中方便迅速地获取相关病症的原因、症状、治疗等信息，也可以在线咨询医生、在线挂号，满足了中国社会对此的巨大需求。这也是大数据极具前景的应用领域，百度在国内走在了前列。

轻应用

“轻应用”，即大数据在移动产业的应用产品。百度致力于在移动互联网上发展可以提供方便的轻应用，努力实现让普通用户获取方便、使用方便的目标，既开发自己的数据源，也让数据成为用户随时可以获取的“价值”。

☆开放的大数据实验室

百度公司建立了大数据实验室，来为产品研发和推广服务，而且已经将云计算和大数据能力开放给了开发者，降低了开发门槛。这就给了普通的开发者一个巨大的应用舞台，让他们能够站在巨人的肩膀上，促进大数据技术的共同进步。这个开放的平台，可以为开发者提供开发、测试、运营等全方位的服务，很多工具和服务都是依托于百度的大数据能力。

现在，大数据在中国已经迈入了蓬勃发展的阶段，像阿里巴巴、百度、360和腾讯等千亿公司开始成长为中国的大数据巨头。在它们的带动下，中小企业也应该量体裁衣，结合自身的情况迅速跟进，来共同创建中国的大数据产业链，协同合作，在政府的引导和大企业的带动下，不断地探索与创新。

CHAPTER 5 大数据与技术变革

告别小数据时代

我们依托某一个独立的数据点产生的直觉来分析、判断问题，就是很典型的“小数据时代”的技术做法。但是，这种靠某一个点产生的直觉和数据判断，只能够解决日常问题，面对复杂的信息流时，它往往会使我们在归类和决策时误入歧途，从而产生一系列的错误——

某个信息“点”可以由点及面，对普遍规律做出模糊预测吗？

个案是否具有代表性和广泛性？

今天发生的，明天是否还会发生？

如果缺乏对于连续数据和多领域数据的宏观统计，以及对相关技术的研发和使用，人们就可能会被最新和最近的“数据点”搞得眼花缭乱，而失去了对于大局的整体把握。

大数据技术与小数据技术恰恰相反，它更多是一种宏观的技术思维，是让我们从“盘子里”跳出来，以更宽阔的视野寻找答案的动力，是帮助我们从各种类型的数据中综合而且快速获得有价值信息的能力。

就像操作系统一样。如果说小数据是安卓（只能用于手机），大数据就是XP。它承载更多，速度更快，分析更准，容量更多元，且能引发一场技术性的变革。

在技术准备上，与小数据的单一相比，大数据也更为广泛，几乎穷尽现今的一切互联网技术，包括大规模并行处理（MPP）数据库，数据挖掘电网，分布式文件系统，分布式数据库，云计算平台，互联网和可扩展的存储系统等。

拿我们的生活来说，假如你使用大数据技术来管理自己的日常生活，那么就不能只靠一个个简单的没有关联的Excel文档进行统计，而应该使它们互动起来，建立一个综合数据库，分类分析和总结，进而才能改进我们的“生活管理”。

消费数据：

在生活中我们对消费的统计是很重要的一件事，我相信几乎人人都有一本家庭账目。但对消费数据怎么统计和分析，使用什么样的工具和方法，最后的结果是大不相同的，甚至会出现截然相反、冰火两重天的对比。

像小数据的统计方法，无非是罗列式的，1月花了多少，2月花了多少，挨月记下来，年底一汇总才发现：呀，这一年我竟然花了7万元在没必要的事项上，超过了自己的计划，怎么办？只能接受现实。但用这种方法，第二年往往还是这样，解决不了问题。

如果换一种思维，我们可以给自己引入大数据的统计技术。比如根据消费计划，分别列出不同的表格，重新记账、分类和分析：哪些钱是该花的？哪些是不该花的？哪些钱的消费属于一时冲动？哪些属于我们被商家忽悠的？

重要的不是统计，而是寻找原因。我们革新思维，重新利用技术。技术并没有划时代的突进，只是使用的功能变了。因此，我们就能够从中找出自己在消费时乱花钱的诱因，再视具体的情况做出改变，并且要每周统计、每周分析，实行消费支出预算制，才能达到理性消费的目标。

时间数据：

大数据在时间管理上也有它高效的应用。怎样将你有限的时间和精力进行合理的分配？如何让自己拥有最高效率的时间应用？小数据思维是确定工作计划然后按计划去做就可以了，严格地执行计划表，是小数据时代的时间管理原则。但大数据时代则不同，你首先需要确定目标，然后将你要做的事情，根据重要程度分清顺序，再罗列计划和执行方案。在时间的监督过程中，你还需要随时调整不同事项的紧急程度，灵活地变革计划，以让自己的时间达到最高效的使用。

工作数据：

在小数据时代，工作数据是一本流水账，你看到的都是枯燥的记录。大数据则让它变成了一座工作数据的储藏宝库。而且，这不仅仅是一个通过数据进行工作记录的过程，更是我们不断总结和认识自己的成长历程。

大数据不会为你提供最终答案，它记录的一切都只是为了让你参考——你能方便地分析过去，总结现在，收获经验，以便获得更好的工作方法以及工作的方向。

我的一位朋友研究数据技术的各类应用已有十几年的时间，他现在最大的感受就是："技术的变革虽已开始，但许多人还没有认识到，自己应该尽快忘记小数据的技术思维，马上投入大数据的技术世界，才能抓住未来。这表明技术的重组其实是在构建一个更高的平台，它需要我们思维的进步，然后才能心安理得地享受数据技术的新功能。比如，十年前我们搞街头调查的时候，只需要一次抽样调查就可以了，这是小数据；但是现在，却需要在全球几十个国家的上百座城市同时进行一次调查，然后汇总数据进行实时分析，并迅速得出结论，这就是大数据。很显然，我不确定从业者都已紧跟潮流。"

现在，我们建立"大数据技术"的动力有哪些？

第一个动力是，我们必须明确："我们是否真的需要大量的数据？"

这是一个“大数据到底为何存在”的问题。假如你不能先解决这个问题，你就会盲目地为了拥有大数据而去变革技术，付出无效的代价。有的人对大数据满怀期待，希望能够发现过去没有认识到的东西，收获惊喜的结果，最后却发现“这些东西我们根本用不上”或者“这不过是已有的事实”而已。

就像有的公司，为了系统开发投入了几千万元，信誓旦旦要迈入大数据的门槛，告别小数据时代，最终得出的不过是证明了资深员工的“经验”的结论，这就太让人难以接受了。只有“需求”才是最大的动力，这是技术进步的主要推动力。如果你没有需求，那么小数据技术也挺好。就像你使用手机如果只是为了接听电话和发一发短信，从来不会上网和视频通话，为何还要花很多钱购买苹果手机呢？

第二个动力与维护数据的需求有关：由谁来维护大量的数据，才能保障数据的质量呢？

换句话说，我们的技术（技术人员）能否保证收集和整理到高质量的数据？

比如，一家公司的部门主管每个月都会收到某客户的宣传资料，但收件人的头衔并不是“部门总监”，而是他在前一段时间兼任公司市场部经理时的头衔。虽然这不是什么大问题，他也仍然会按时收到这些资料，但他还是提出，希望对方改变一下头衔。

客户经理当场道歉，并表示回去会马上进行修改。但到次月，这位主管再次收到资料时，发现收件人的头衔没有任何更改，仍然写着几个大大的字：市场部经理。他非常失望，然后决定中止与该客户的合作。

问题出在哪里？仅仅是对方没有重视这一个细节吗？当然不是，归根结底，该客户公司缺乏维护顾客数据库的意识，在收集和整理数据时，工作充满了疏忽，无法保持高质量数据的实时性。小数据时代不必在乎这些，但在大数据时代，“企业外部”的数据是否最新、是否精确，都是一件极为重要的事情。

如果你的人员收集而来的数据出处不明，或存在严重错误，那么数据将毫无意义；如果这些数据不实时维护与更新，不是最新的数据，也不会产生任何价值。

第三个动力是我们的工作激情与事业规划。

具体地讲，就是企业的战略规划与员工的事业目标是否完美地结合了起来。如果没有，那么员工的工作激情就会成为问题。当你希望他们与公司一起迈进大数据时代、布置大数据技术时，他们的思维仍然是“小数据时代”的，在相当长的一段时间内，使用的也必然还是小数据技术。

这也告诉我们，在大数据的技术革新中，人的因素永远是最重要的。企业的方向是努力培养我们的数据科学家，同时提升现有人员的分析数据的能力，提高他们的激情，增加他们分析和利用数据的意识。假如我们的每一名员工都十分擅长“数据”，也对数据有极高的敏感度，经常可以自主地通过数据考虑事情并进行判断，你的公司一定能够强大起来，也必然会强大起来。

重要的是，数据为工作带来成效，也能由此让员工的工作动力更加充足。

这三点对于大数据的技术应用非常重要。中国人对于大数据已经期待很久了，我们也开始在各类新闻节目和财经频道中看到它的影子，但要想从小数据的技术习惯中彻底摆脱出来，让大数据真正成长壮大，还需要很多的努力。

数据服务产业链

管理学大师德鲁克说：“当今企业间的竞争，不再仅仅是产品的竞争，更是商业模式的竞争。”

大数据要想落地，必须有三个条件：一是丰富的数据源，二是强大的数据挖掘和数据分析能力，三是建立完善的数据服务产业链，也就是商业模式。商业模式指导着公司如何赚取剩余价值，因此确立公司在产业链和价值链中的位置，至关重要。

现在，在IT领域，已经逐步降低了分析技术的门槛。很多企业因为数据源匮乏，因此在各自的大数据战略上纷纷受挫。它们感到迷茫：“都说大数据是机遇，但机遇在哪儿？”它们也由此迷失了自己在数据服务产业链中的位置，反而感觉不如以前，因此萌生出还不如小数据模式的想法。

企业要想在大数据时代领先，必须获取更多的数据，并且明确自己的商业模式，在价值链中如何选取上下游合作伙伴以及怎样与客户达成交易、为客户提供价值。要知道，这是大数据的基础，更是大数据战略成败的核心。许多企业的迷失，恰恰是在新的产业链条中对于商业模式和产业分工的迷茫。

在大数据时代，共有三种大数据公司活跃在大数据产业链上：

1.数据拥有者：基于数据本身的公司。拥有大量数据，但是不具有数据分析的能力。

2.技术提供者：基于技术的公司。例如技术供应商或者数据分析公司等。

3.服务提供者：基于思维的公司，也就是挖掘数据价值的大数据应用公司。

扮演着不同的产业链的角色，就具有不同的盈利模式。我们可以对大数据的商业模式做一下梳理和细分，以供读者参考。

☆“数据拥有者”的商业模式

数据拥有者的公司共有三类：

1.对大数据的重复利用是其发展的原动力，其中大数据是它们业务的核心，这种公司具有很强大的大数据技术能力，同时具有三种产业链角色：数据+技术+服务。多数时候，它们公司的技术用于自身的运作。例如谷歌、亚马逊、百度、阿里巴巴等这些世界级的互联网企业。

2.大数据是为提高自己公司的生产效率、增加业务收入或者创造新的收入提供基础的，并非厂商的主流业务。例如运营商、银行等，目前运营商本身并不通过数据的重复利用来盈利，其主要业务是通信设备提

供的各种网络语音和数据业务。

3.数据中间商。此类公司从各种地方搜集数据进行整合，然后再提取有用的信息进行利用，把这些高价值的数据提供给需要的公司，但是它们本身不具有创造数据的能力。比如一些调查公司等。

这些数据拥有者的商业模式有：

2B：提供数据分析的结果，主要面向企业或者政府部门。例如Inrix公司出售完整的交通状况的模式图给交通规划部门、物流公司、GPS生产商等。

2C：提供基于数据分析结果的服务，主要面向个人。例如Inrix公司为用户提供免费的交通信息，但是这是一个免费的智能手机应用程序，用户可以自行下载，该公司自己却可以得到同步的数据。

2D：把数据或信息作为资产直接进行销售，并且构建一个数据资产分享和交易平台，这是一种全新的商业模式。例如推特通过两个独立的公司把它的数据授权给别人使用；VISA公司通过收集和分析210个国家的15亿信用卡用户的650亿条交易记录来预测商业发展和客户的消费趋势，最后把这些分析结果卖给其他的公司。

☆“技术提供者”的商业模式

技术提供者们目前主流的商业模式是2B，其中有4种类型：

1.提供单点技术为主。例如，Teradata公司为沃尔玛这个大型零售商提供大数据分析技术。

2.提供整体解决方案，其中以IT厂商为主。例如，著名的IBM公司提供了一套完整的大数据解决方案；中国华为的大数据解决方案则是基于IT基础设施领域在存储和计算的优势。

3.大数据空间出租。通过出租一个虚拟空间，从简单的文件存储，逐步扩展到数据聚合平台，在大数据计算基础设施上与云端有机结合。例如，腾讯公司的“开放云”战略中，小企业也有机会在大数据领域创新业务，为大数据创业者提供了廉价的数据基础设施。

4.提供E2E在线大数据技术或者解决方案，即Bigdata as a service。简言之，这是一种新的商业模式。例如，RJMetrics公司有一款软件，客户只需在软件端输入特定数据，该公司便会在7日内优化数据，将这些信息备份到安全的服务器上之后，以清晰简洁的界面将数据分析结果反馈给客户。该软件的定价只有每月500美元，却能够为电商提供快捷的商业智能在线服务。另外，GoodData公司则面向商业用户和IT企业高管提供数据存储、性能报告、数据分析等工具，其中，所有商业智能分析所需的数据都将在云端进行。

技术提供者也有2C商业模式，但是目前还比较少，与云端结合后却

会有很大的空间，这在未来将会是一种巨大的趋势。例如，有些公司面向个人的家庭帐单、家庭耗能节能等或者面向个人数据的一些大数据解决方案。假如你有需求，并有支付成本的能力，你可以立即得到这种便捷的服务。

☆“服务提供者”的商业模式

大数据的服务提供者有两种，一种是应用服务提供者，另一种是咨询服务提供者。

1.应用服务提供者是对外提供服务的，它是基于大数据技术的一项服务。

商业模式——

2B商业模式：提供数据分析结果的服务，主要面向企业或者公共政府部门。例如前面提过的Inrix公司。

2C商业模式：提供基于数据分析的服务，主要面向个人。例如，Flight caster公司和FlyOnTime.us公司通过分析过去十年里每个航班的情况并且将其与天气情况进行匹配，以此预测航班是否会晚点。

2.咨询服务提供者是头脑风暴的大赢家，他们主要提供技术服务支持、技术(方法、商业等)咨询，或者为企业提供某种咨询服务，类似数据科学家。

商业模式——

2B商业模式：他们通过大量数据支持，利用数据挖掘技术帮助客户开拓精准营销，对数据进行挖掘分析后预测相关主体的行为。他们的收入来自客户增值部分的分成。例如德国咨询公司GFK提供基于地点的人员流动的数据，以时间为维度，分析特定区域的人员人口统计数据(性别、年龄)和行动等数据，主要面向零售商、政府部门、公共机构等。这类企业成长非常快，擅长数据挖掘分析技术，帮助一些数据大户如银行、运营商等开展新的业务。

目前，数据服务产业链上真正的大数据玩家，应该是例如谷歌之类的公司，通过重复利用数据以获得利益。谷歌成功地建立了“网页搜索+广告”的商业模式，其所有的业务都是构建在大数据之上的。

因此，谷歌是大数据最大的玩家。2012年，它的总营收达到501.75亿美元，利润107.4亿美元，其九成利润来自广告。有咨询公司预测，2017年全球大数据技术的市场空间约500亿美金，约等于谷歌2012年的总营收。这其中，既包括了技术，也包括了大数据工具和相应的服务。

由此观之，在大数据时代的未来，获利最大者将是“数据为王”或者“数据驱动”的业务内涵和模式，发展大数据并挖掘大数据的新价值是其不可不为的原动力。中国也必须尽快扶植相关的产业和公司，以抗衡

谷歌这样的跨国巨头，并最终实现赶超。

技术支持与发展

大数据不是口号，而是技术，同时也是技术的整合。大数据的到来，已经成为现实生活中无法逃避的挑战。无论如何，大数据已经成为新一轮技术变革的最强音。关于模式的思考，关于安全的质疑，关于应用的探索，我们必须静下心来看大数据，实实在在了解大数据目前仍然需要解决的问题。

国计民生、商业创新无不与大数据相关，大数据渐渐向人们展现了它为学术、工业和政府带来的巨大机遇。每当我们要做出决策的时候，大数据就无处不在。无论如何，我们都必须直面大数据时代的到来。

大数据给中国带来的巨大挑战，首先是三个重要的技术问题。

☆如何利用信息技术等手段处理非结构化和半结构化数据

大数据的一个重要特点就是数据分散。大数据中，85%都是非结构化的数据，结构化数据只占15%左右。大数据的另一个特点就是不确定性，表现在高维、多变和强随机性等方面。有90%的数据来自开源数据，其余的被存储在数据库中。而大数据则大量存在于社交网络、互联网和电子商务等领域。

值得注意的是，大数据刺激了大量的研究问题。但是大数据每一种表示形式都仅呈现数据本身的侧面表现，并非全貌。比如图像，如何把它转化成多维数据表、面向对象的数据模型或者直接基于图像的数据模型？

如果把通过数据挖掘提取“粗糙知识”的过程称为“一次挖掘”过程，那么将粗糙知识与被量化后的主观知识相结合而产生“智能知识”的过程就叫作“二次挖掘”。这些结构化的粗糙知识可以被主观知识加工处理并转化，生成半结构化和非结构化的智能知识，这也正是基于大数据的数据挖掘所产生的结构化的粗糙知识的一些新特征。

由于大数据所具有的半结构化和非结构化的特点，寻求“智能知识”也就反映了大数据研究的核心价值。非结构化和半结构化数据的个体表现、一般性特征和基本原理尚不清晰，要想实现从“一次挖掘”到“二次挖掘”这样类似事物量到质的飞跃，还必须通过包括数学、经济学、社会学、计算机科学和管理科学在内的多学科交叉来研究和讨论。这些都需要给定一种半结构化或非结构化数据，包括具体的经验、常识、本能、情境知识和用户偏好。

☆如何探索大数据复杂性、不确定性特征描述的刻画方法及大数据的系统建模

大数据的复杂形式导致许多对“粗糙知识”的度量和评估显得尤为重

要。这一问题的突破是实现大数据知识发现的前提和关键。这里，人机交互将起到至关重要的作用。管理科学，尤其是基于最优化的理论将在发展大数据知识发现的一般性方法和规律性中发挥重要的作用。

从短期而言，学术界鼓励发展半结构化、非结构化数据之间的转化原则，以支持大数据的交叉工业应用。从长远角度来看，可以将已知的最优化、数据包络分析、期望理论、管理科学中的效用理论应用到“二次挖掘”过程中，研究如何将主观知识融合到数据挖掘产生的粗糙知识中。大数据的个体复杂性和随机性所带来的挑战将促使大数据数学结构的形成，从而导致大数据统一理论的完备。

☆数据异构性与决策异构性的关系对大数据知识发现与管理决策的影响

在大数据环境下，管理决策面临着两个“异构性”问题：“决策异构性”和“数据异构性”。大数据已经改变了传统的管理决策结构的模式。决策结构的变化要求人们去探讨如何为支持更高层次的决策而去做“二次挖掘”。探索大数据环境下决策结构的改变对管理决策结构的影响会成为一个公开的科研问题。寻找大数据的科学模式将带来对研究大数据之美的一般性方法的探究，已知的数据挖掘方法将成为大数据挖掘的工具。

无论大数据带来了哪种数据异构性，大数据中的“粗糙知识”仍可被看作“一次挖掘”的范畴。由于大数据本身的复杂性，这一问题无疑是一个重要的科研课题，传统的管理决策模式取决于对业务知识的学习和日益积累的实践经验，而管理决策又是以数据分析为基础的。大数据是一种具有隐藏法则的人造自然，如果我们找到了将非结构化、半结构化数据转化成结构化数据的方法，通过寻找“二次挖掘”产生的“智能知识”来作为数据异构性和决策异构性之间的桥梁，那么我们将能够很好地应对传统的数据挖掘理论和技术提出的新挑战。尽管这样的探索十分困难，但是研究大数据，是十分必要的。

除此之外，还有一些数据科学的问题，以上也仅仅是研究大数据挑战的一个起点。在未来，相关的问题都可以得到很好地解决。

自从人类进入到信息化时代以来，我们不断产生大量的数据，加之物联网、移动互联网应用的大规模爆发，大量新数据以每年50%的速度在增长，或者说每两年就要翻一番多。数据已经渗透到每一个行业和服务职能领域，随着互联网技术的不断发展，数据本身就是资产，这一点在业界已经形成共识。

人们对于海量数据的运用将预示着新一波生产率增长和消费者盈余浪潮的到来。在云计算时代，人类通过对海量大数据的高效分析获得商

业以及社会价值。云时代的到来、移动终端普及使得数据创造的主体由企业逐渐转向个体，而个体所产生的绝大部分数据为图片、文档、视频等非结构化数据。随着云计算技术的快速普及，人类社会正在步入一个被互联网和通讯技术引爆的大数据时代。大数据技术在中国的发展前景是光明的，前提是我们能够提升和扩充自己的技术王国，建设美好的蓝图。

全球技术研究和咨询公司Gartner将大数据技术列入2012年对众多公司和组织机构具有战略意义的十大技术与趋势之一。Gartner在其新兴技术成熟度曲线中将大数据技术视为转型技术，这意味着大数据技术将在未来3～5年内进入主流。中国也不会落后，“云基地”作为国内最早根植在云计算技术及商业模式的领先者，也一直积极关注大数据带来的发展机遇。

从战略到战术层面，从理念到技术层面，中国都已开始自我的进化，更加适应这个新的时代。中国经过了几十年的积累，让不断产生的海量数据正在成为虚拟世界取之不尽的能源，而它们还远未被开发。

信息化技术的普及使得中国企业更多的办公流程通过网络得以实现，由此产生的数据也以非结构化数据为主。而其他领域的研究，如云计算、下一代分析、内存计算等也都与大数据的研究相辅相成。

我们尚无法确定万物是否皆有数据，但是至少已经推开了这样一扇大门：以理性的态度思考大数据，共同保持着持续变革的动力，主动地拥抱这种变化。早在2012年时，非结构化数据就已达到了互联网整个数据量的75%以上，用于提取智慧的大数据，往往是这些非结构化数据。而现在，这个比例已变得更大，我们也拥有了足够的技术支持。换句话说，中国的大数据技术的积累已到达了突破阶段。

“脚印追踪”——个性化的数据推荐系统

如果你在自己经常网购的网站看到“猜你喜欢”之类的东西是那么符合自己的要求，不要惊讶，因为我们每一个人都已经步入了大数据时代。你可以想象一下，也许在未来，每天打开电脑，它会自动把你所有的需求列一个清单，你只需要坐在舒服的沙发上点几个确定选项，就可以轻轻松松地搞定一切了。

不要以为这种情形只发生在科幻电影中。商家只有在满足大众的需求时才能够卖出商品,而这一切都是在满足大众的个性化需求的基础之上实现的。

2011年9月，淘宝公司发起了用户定制电视的活动，2天内1万台订制电视就被抢光。在该活动中，用户可以选择电视的各种属性，包括尺寸、边框、颜色等，厂商根据用户的订制内容生产电视产品，再送货到客户的家中。

从这个具有代表性的案例中我们就可以发现，未来的商业模式正在发生着质的变化，它通过满足个性化需求来提升商业运转的效率，在为消费者提供更好服务的同时，获得更多的利润。

☆“猜你喜欢”的由来

网购中随处可见的“猜你喜欢”是怎么来的呢？事实上，这种推荐方式来源于亚马逊的技术创新。

亚马逊公司的内容起初都是由人工完成的，他们聘请了一个由20人组成的书评团队在网页上推荐有意思的新书。但是随着在亚马逊上架的图书越来越多，这样的人工操作自然越来越显得乏力低效了。

后来，亚马逊的总裁贝索斯决定尝试更有创造性的做法，根据用户的习惯来为其推荐商品。但若想实现个性化推荐，必须要将不同用户进行比较再找到用户之间的关联。但是，面对庞大的数据，这种推荐系统算法烦琐，结果也是不尽如人意。

亚马逊当时的技术人员格雷格·林登思考之后，想出了一个解决方法：其实，根本没有必要将不同的用户进行比较，我们只需要找到产品之间的关联。这样的推荐方式可以提前分析产品之间的关系，所以推荐速度非常快，适用于不同产品，甚至可以跨界推荐商品。

林登说：“书评团队被打败、被解散，我对此感到非常难过。但是，数据没有说谎，人工评论的成本是非常高的。”

他将书评家带来的销售量和推荐系统产生的营销业绩进行了比较，发现推荐系统带来的商品销量远远高于书评家，这个销量比较数据直接影响了亚马逊解散书评组，而由推荐系统取代他们来推荐更可能受用户

欢迎的产品。

在亚马逊的带领下，越来越多的公司开始使用这种个性化推荐系统，迅速推动了电子商务的发展。而这种基于海量数据的推荐，也是大数据早期运用的一种形式。

☆大数据是实现个性化的基础

事实上，要实现个性化的商业模式，充足的数据是其必不可少的基础。没有海量的数据，个性化也无从谈起；没有海量的数据，我们甚至连小部分用户的个性也很难总结，更不用说多数用户。

不知道你是否听过“啤酒和尿布”的经典故事。在超市里，尿布要摆在啤酒旁边才能卖得好，这也正是在深入分析大众需求的基础上得出的结论。这样的“规律”安静地隐藏在数据中，它一言不发，只等人们自己发现。我们总是需要深挖，才能让它们浮出水面。

然而大数据相对于传统的数据挖掘更进一步。大数据到底有多大？一组名为“互联网上一天”的数据告诉我们，一天之中，互联网产生的全部内容可以刻满1.68亿张DVD；发出的邮件有2940亿封之多（相当于美国两年的纸质信件数量）；发出的社区帖子达200万个（相当于《时代》杂志770年的文字量）；卖出的手机为37.8万台，高于全球每天出生的婴儿数量37.1万……大数据具有如下特点：数据量大、数据种类多、数据之间有潜在关联、速度快、时效高。

无所不在的数据、无处不在的网络和大规模分布式的存储和运算能力（云计算），忠实地记录了我们的衣、食、住、行及社交状态。现在，人类一天创造的数据相当于2000年一年的数据量。

你是否每天会在微博、微信、人人网之类的网站上发布信息？一分钟内，微博上新发的微博超过10万；社交网站Facebook的浏览量超过600万……整个互联网的用户和所有的商品本身就是一个足够大的数据空间，加上空间、时间、天气等等潜在相关因素，想要知道每个用户的喜好，所需要的数据量是巨大的。数据越多，对于用户的理解越精准，但同时分析的难度也就越大。

☆互联网大数据处理的技术挑战

事实上，当你仍然在把微博等社交平台当作抒情或者发议论的工具时，华尔街的敛财高手们却正在挖掘这些互联网的“数据财富”，先人一步用其预判市场走势，而且取得了不俗的收益。

然而，处理互联网大数据充满挑战，正如上文所提到的，数据如此之多，如此之庞杂，如何才能够有效地找到数据之间的关联，如何才能充分利用大数据以期实现个性化的需要，都是我们每一个人需要思考的问题。

我们首先要明确的是，处理大数据需要具备哪些能力。为使消费数据的速度赶超生成数据的速度，拥有足够的计算资源是必要条件。而大数据处理的核心能力为具有高水平的计算框架、稳定的程序设计以及精准的算法。而这些能力则需要专业的计算机技术人才来实现。

其次便是时效性。用户生成数据的速度是非常之快的。如何才能及时感知到这些有效数据，在用户下一次操作前做出有效的响应，最终给用户带来便捷？这样的时效性要求计算机系统能够以数据流的方式来运转，最终导致系统采用与传统批量大数据处理截然不同的技术方案。

最后，为了更大程度地满足个性化需求，还必须具有足够强大的定制能力。尽管单个用户的定制需求可能很小，但用户数量巨大，定制需求迥异，如何才能够及时有效地满足每一位用户的需求呢？这就需要有像数据库SQL语言（结构化查询语言）那样给用户足够多的自由，使再小的需求通过简单的操作就能满足。这样的定制能力要在数据的存储、运算、查询、展现等多方面都有体现。

☆阿里云的解决之道——云推荐

不论是收集大数据的计算和存储能力，还是处理个性化问题所需要的实时计算和算法技术，对于网站站长和开发者而言都是不容易快速得到解决的问题。听说过“云”的概念吗？它相当于把每一位用户生成的数据内容存储到一个大的存储器中，再根据用户的需求从“云端”下载，比如现在非常流行的“云电视”、“云储存”、“云服务”等。阿里云正试图通过云端服务来降低个性化服务的门槛，使更多网站站长和开发者能够低成本享有自己的个性化服务，其中云推荐便是一个典型。

什么是云推荐呢？举一个最简单的例子来说，如果某网站是介绍美食菜谱的，用户在浏览某种菜或者是汤的制作方法时，如果能够有些相关菜谱的推荐，那么便可以让用户在网站内停留更多时间，访问更多内容。

云推荐又该如何实现呢？事实上，有多种方式可以找到用户感兴趣的内容：

第一，从用户访问日志里面发现。每一位用户都会产生浏览记录，云推荐通过对用户浏览记录数据的科学分析，推荐一系列的相关内容。

第二，可以把网站里面其他热门的菜谱推荐出来，对不同网站相同的内容进行整合与链接。

第三，寻找不同种类的内容，假如用户在浏览某类汤的做法，那么可以推荐一些某类饭菜的做法。

然而，要实现这样的推荐，传统的做法需要大量的人工编辑工作。既不能做到即时，也很难保证好的效果。一个精准的推荐模型，必须对

该方法本身的整体效果以及用户对各种推荐方法结果的偏好作出一个综合的评估，这样才能找到适合每一个用户的精准推荐模型，最终让用户享受到推荐展位“千人千面”的个性化服务。

那么，普通人可以使用云推荐服务吗？

可以的。如果你也想尝试一下，只需在云推荐网站注册申请，得到一个十位应用ID，如“1000001234”，并将系统生成的代码内嵌在网页代码中便可以得到个性化推荐结果。这个过程一般一分钟即可完成。

随后的事情，当然就要交给云端系统了。它会开始对网站进行深度分析，还会持续根据展现的点击效果自动调整推荐方法的模型和权重。

在云推荐的管理界面里，网站开发人员可以定制推荐位置大小、推荐内容条目数、URL范围、展现形式等参数。网站站长还能看到推荐展位的点击情况，并根据建议适当调整推荐位置参数以改善效果。

如果你是专业的网站运营和管理人员，那么你要知道，云推荐服务还针对主流建站工具Wordpress等提供插件支持。开发人员在安装插件之后，即可在工具管理界面来操作并管理云推荐的各项功能。根据后台统计，网站启用云推荐后的整体流量会提升10%。这样的个性化服务让人感觉就像是钱存银行能拿到利息一样，是大数据魅力的展现。相信随着数据的不断积累及用户数量的累积，个性化服务在大数据时代能给人带来的远不止10%流量提升这样的惊喜！

现在，你也许能够理解云推荐背后的奥秘了，它的基础仍旧是每一位用户生成的数据，你浏览的什么网站，在网上发布了什么东西，或者点了某篇文章，都是海量数据的基础。正如海纳百川的道理一样，数以万计的网民产生的数据汇聚到一起，就成为“大数据”。它的本质是如此透明，但又是这么巨大。

只要通过专业的分析软件，把大数据分析、整合、利用，最后就会成为每一位用户所看到的“猜你喜欢”。这个时候，想必你已经不再感到惊讶了。秘密就是这么简单！

☆个性化真的安全吗？

每一件事情有好的一面，就有不好的一面。大数据也是如此，无论它发展出多么奇妙的应用和提供多么方便的功能，它都有需要我们规避与纠正的弊端。

让你印象最深刻的一定是隐私。大数据在大大方便了我们的生活的同时，也在严重威胁着我们的个人隐私。一方面，我们不得不使用网络，但另一方面，我们又害怕自己的隐私被毫不遮掩地暴露。“棱镜门”事件让我们重新思考大数据下的安全问题。随着“棱镜门”的发酵，大家惊讶地发现，美国的网络监控已经在大数据盛行的今天走得那么

远，将世界各国都抛在了后面。

同时还有根服务器的问题。在1969年，美国西南部的4所大学——加州大学洛杉矶分校、斯坦福大学研究院、加州大学、犹他州大学的4台主要计算机连接起来，这就是最早的互联网。目前，中国互联网用户已突破5亿，全球排名第一，但主要用来管理互联网主目录的根服务器全世界共13台，1台主根服务器仍然在美国，其余12台辅根服务器9台在美国，没有1台在中国。

美国的互联网用户数量还不到我国的一半，但网络主机数量是中国的28倍。在美国控制了多数根服务器的前提下，中国的网络安全堪忧。这也是个性化存在风险的决定性因素之一，因为我们无法控制现状，就像一个人被掐住了脖子一样。

前有传媒大亨默多克的《太阳报》因为窃听丑闻而关闭，后有斯诺登事件的发生，这些都在考验着我们着实堪忧的网络环境。所以，每一个人心里都有一个问号，大数据环境下的网络安全吗？我们的隐私还是隐私吗？如何才能安全地使用互联网，在大数据的使用和用户隐私之间找到一个平衡点，是我们需要进一步思考的问题。

CHAPTER 6 大数据与思维变革

思维数据化——赢在大脑

“怎样才能赢？”这是一个古老的问题，在数千年的文明史中有无数的答案。有人会说：有力量就能赢；也有人说：有财富才能赢。但至少在今天，我的答案是：谁具备了大数据思维，谁就能成为未来最大的赢家。

从目前来看，当大数据时代到来时，任何一家公司的竞争力都可以划分为三种类型。第一种是大数据本身；第二种是与大数据相关的技术；第三种是大数据思维。这三种竞争力当然都是不可替代的，也是缺一不可的，但其中最为关键的，就是把数据与思维结合起来的部分。数据可以被复制，技术也可以被超越，只有思维难以被窃取。拥有领先思维的大数据玩家，最有资格发动一场胜算极大的战争，或者占据最大份额的市场，形成自己坚不可摧的竞争力。

同时我也发现，具备大数据思维优势的公司往往是那些新兴的创业型公司，它们在一个全新的领域内崛起，而且它们的创始人大都具备大数据思维能力和大数据技术，能够及早地发现某特定商业领域中大数据的应用价值，并且做到第一时间把自己的理想付诸实施。在别人进入之前，它们就已完成了垄断。

做企业是这样，生活难道不是这样吗？它对个人也是适用的。你在公司给人打工，怎么让领导发现你的价值？如何让上司觉得你比同事更强？大凡竞争中的赢家，他在今天一定要懂得用数据作为衡量标准，用数据来看待事物。当你拥有这种思维模式时，你会发现这个世界变得完全不一样了，你将比以前更加清醒，也更加理性。

大数据时代的到来，不仅是技术的更新，它同时标志着我们处理信息方式的变化，我们思考问题的模式的升级，我们思维深度的掘进，也是我们智能的进化。随着时间的推移，大数据将会彻底地改变人们思考这个世界的方式。

我们将懂得如何从大量的信息中学习到从较少量的信息中无法获取的东西；

我们将利用越来越多的数据来理解事情和作出决定；

我们将发现许多东西是随机的而不是确定的；

我们将认识到事物之间相关性与因果性的区别。

当然，我们更会强烈地觉察到，今天的世界已经是一个智力博弈的时代，也是属于思维为王的新世纪。随着大数据处理能力的提高，大数据产业对于工作和生活各个层面的渗透，人类的观察力、记忆力、想象力、分析判断能力、思维能力、应变能力也将会得到质的提升。

你会逐渐而且惊奇地发现，不但自己更加理性，城市也在变得智能化。未来的世界将彻底变成一个大脑为王的竞技场，金钱和权力都将成为大脑的奴仆。人类的大脑创造数据，也将支配数据。

其实早在五六年前，行业领先者们就已预言：大数据的到来将引发一场新的“智慧革命”。我们可以从海量、复杂、实时的大数据中发现知识，提升智能，为社会创造更大的价值。所以，尽管存在这样或那样的不足，但大数据时代一定是美好的时代，因为数据化正在可控的范围内让我们的生活更美好，让我们的工作更方便，让我们的未来更清晰，也让人类看到了改变世界整体结构的希望，让它逐步具备“智慧”特征，从而通过数据这个工具，实现人与自然的沟通，互相之间进行智慧与理性的交流。

那么，到这时候，我们的学习、工作、生活、娱乐以及交通、医疗、能源利用方式等等都将随之改变。我们可以改变自己的头脑，从海量数据中获取所必需的工具和技能；可以提升自己的智慧，以大数据的思维重塑自己的人生战略，增强竞争力。

中国的大数据逻辑：因果关系 > 相关关系

在今天的时代，我们需要更换思维，以全新的头脑和运算逻辑来看待每一件事物，才能从中发现新的东西，释放更多的生产力。大数据就为我们提供了一种可以全视角观察世界的新方式。

☆重要区别——从因果性的分析到相关性的分析

国外的大数据思维认为——如同欧美的大数据专家们在采访中所讲的，一项重要的思维转换就是从传统的因果性分析向相关性分析转换，只有以相关性为主，才能真正体现大数据的思维特点。他们认为，大数据的出现有一点是常被人忽视的，那就是悄悄地改变了人们过往普遍追求的因果性的思维逻辑。

美国大数据专家罗伯特说：“大数据主要从相关性着手，而不是因果关系，这从本质上改变了传统数据的开采模式。”

比如，他举例说，谷歌的研究人员在2009年发表了一篇论文，成功地预测到了季节性流感的爆发，引起了医学界的轰动。研究人员对2003年到2008年之间的最频繁搜索词条（多达5000万）进行了非常全面的分析，希望能够发现某些搜索词条的地理位置的特征——是否和美国流感疾病预防和控制中心发布的数据相关。

罗伯特说：“疾病预防控制中心会定期跟踪全美各地的医院以及私人诊所的病患，然后汇总和发布相关的信息，但往往会滞后一两个礼拜，这是必然的，这些信息的人工整理需要一定的时间，但谷歌的大数据能发现实时的趋势，这些词条都是实时的，有对应的时间和地点的记录。”

最后，谷歌公司将得出的预测与疾病预防控制中心记录的最近两年内的实际流感病例进行对比后发现，大数据处理结果找到了45条检索词条的组合，通过合适的数据模型计算后，通过相关性的预测得出的结论，与官方数据的重叠性高达97%，这表明通过相关性分析是可以解决此类预测问题的。

欧洲有一家航空公司，它的会员不下数百万。会员的一个重要信息就是邮箱的地址。另外，推特的账号申请也需要一个邮箱地址。通常来说，同一个邮箱地址就意味着航空公司里的会员和推特的会员应该是同一个人。

于是，这家航空公司就做了一个筛选，从中归并出了十万个用户。接下来做什么呢？航空公司请了一家第三方公司的数据部门过来，任务就是看一看这十万名用户会在社交平台上干些什么，比如他们说些什么，关注些什么，以及喜欢介入什么样的话题去转发评论，或者喜欢关

注一些什么样的商业媒体。

这家航空公司的目的在于，研究一下自己需要在社交平台上发起什么样的活动，以及给予什么样的礼品或折扣，才能吸引这十万名会员前来参加，成为公司的贵宾用户，给公司提供利润的增长点。

在这个故事中所涉及的数据虽然足足与十万个人有关，但还算不上海量数据。但它的本质其实就是体现了相关性的价值。航空公司寻求相关性，从而判断自己新的利润增长点在哪里，以及发现潜在的贵宾用户有哪些，然后据此做出高效的决策或采取针对性的营销活动。

☆中国的大数据哲学——因果关系优先

相关性当然十分重要，通过上面的例子我们也已经体会到了它的魔力。通过相关性，我们发现了它对于预测的巨大价值，而其背后则是思维与分析方式的更新。相关性帮助我们从对过去的理解变更为对未来的预测。

但是，因果关系仍然是相关性的逻辑基础。因为数据并不仅是冷冰冰的符号，数据只是事物之间发生联系的代表，而我们每个人都可以将自己作为一个普通个体的因素包容到这个分析体系中，个人的主观的东西会极大地影响体系的方向与决策，这种影响甚至是决定性的。比如人的本身存在的各种因素：风险、意外、热爱、冷酷，甚至是某些错误，都可以在大数据的变化中体现出来，这是相关性无法体现的，必须由因果性加以定义。

“那么，重要的问题来了，因果关系跟相关关系的关系是什么？”

因果关系代表主观性，是人的因素；相关关系代表客观性，是信息的因素。人与信息是结合的，不可分割，也就意味着因果关系与相关关系不可分离。尤其对于中国人来说，我们看到了一个相关性，就会想了解为什么，探究其背后的原因，而不只是商业或市场机会，也不仅仅代表某种现象。

当你开始给出一个假设，建立一个模型，然后去验证这个模型时，这里面就会立刻带入你自己的主观因素，也就是原因。原因即因果性，它决定我们的方向。

这种全新的思维方式和先后顺序是非常重要的。如果你只重相关性，就会因为缺乏因果关系的支撑而背离分析的初衷；如果只重因果性，则会由于忽视了相关关系，而在数据收集时丧失对海量数据和关键信息的把握。

简单优于复杂

在大数据时代，一些烦琐的数据管理流程可能只是一片“云”就能解决的工作。比如一个客户订单通常需要经过ERP、供应链管理、产品数据、库存等多个企业数据管理系统。这很复杂，其流程就像人的思考一样，需要不同的工序来加以实现，最后完成选择，采取行动。

而目前，中国绝大部分企业的数据化管理模式仍旧采用一个个独立分散的系统。正如某位研究大数据的人士所说：“如果能有一个合适的云存储系统，就可以将这些数据整合在一起，做到对企业运行一目了然。”

云存储便是实现这些功能的前提，它让复杂的问题变得简单了，而它也正是大数据思维的集中体现，让一切复杂的问题变得简单。我们在理解大数据时，必须汲取这种宝贵的营养，让它渗入我们新的思维中，用集中化和简单化的思考去解决问题。

让我们想一下巴菲特的忠告：简单胜复杂。巴菲特经常可以把烦琐的投资问题用最简单的逻辑解释清楚，比如他的价值投资哲学。为什么我们大部分人不爱简单爱复杂呢？越是真理越简单，越是谬论也就越复杂。

这个世界上的成功之道表明，凡是真正的成功哲学，它通常不是非常复杂的，而是非常简单的一个系统。不仅在投资理念上，也在做人的哲学上，更在于企业的管理和生产控制上。比如云存储，它能够带来更为简便、精确的效果。

在办公文件的高效、规范管理上，云存储也大有可为。首先需要搭建服务器，然后再由专人进行日常维护，但投入高，操作复杂，专业性强。一项工业设计的多个修改版本，即使通过不同文件名来区分仍可能出现混淆不清甚至丢失的情况，尤其对于工业设计这样的特定行业来说，办公文件的规范管理可能更受重视。企业如果要建立自己的数据中心整合数据，只需根据自己的需要租用中心服务器的一个空间，便能实现云储存和运算服务。

巴菲特说，他的价值投资之道非常成功，却非常简单。多简单呢？简单到“三高”都不需要：一不需要高等数学，二不需要高学历，三不需要高智商。这三种复杂的东西都是无用的，因为都偏离了投资的真相。

他说：“我从来没发现高等数学在投资中有什么作用，只要你懂一些小学算术就足够用了。如果高等数学是必需的，我就得回去送报纸了，我从来没发现在投资中高等数学有什么作用。要想成功地进行投资，你不需要懂得什么专业投资理论。事实上大家最好对这些东西一无所知。

投资并非智力竞赛，智商高的人未必能击败智商低的人。”

老子的《道德经》中有一句话，叫作“道可道，非常道。”又说：“吾言甚易知，甚易行。天下莫能知，莫能行。”讲的都是巴菲特所言的这个道理。有一位工业设计公司负责人对我说：“使用传统依靠文件名来区分的存储模式很容易造成版本的混淆，这其中再加上设计师人员的流动、异地协作设计等因素，经常造成设计成果的丢失。”

因此，工业设计就是一个需长期与文本资料打交道的行业。一项设计成果的最终定型要经过设计师们的反复修改，这其中经历了设计、修改、审核、讨论、再修改、再审核的多个环节后，就产生了许多设计版本。那么要想获得利润还有可能吗？很难，在这种设计思维的主导下，利润就会被压缩到最低。

对任何一个行业、一项事业来讲，要想获得最大的利润，只有两个字：简单。

可以不精确，必须尽量多

在我们从技术层面来萃取或者处理数据的时候，思维的混乱也会发生。其实，混乱的起源和类型本来就是“一团乱麻”。比如，我们在利用Twitter的信息进行情感分析来预测好莱坞票房的时候，就会出现一定的混乱。

在这其中，混乱的表现其实就是格式的不一致。我们要想达到格式一致，就需要在进行数据处理之前仔细地清洗数据，而这在大数据背景下是很难做到的。

为了规模的扩大，我们往往接受适量错误的存在，当然也包括思维的错误。正如技术咨询顾问凯艾尔先生对我说的，有时得到2加2约等于3.9的结果，也很不错了。值得注意的是，错误性并不是大数据本身固有的。它只是我们用来测量、记录和交流数据的工具的一个缺陷。

大数据也不需要进行抽样才能获得最后的结果，以得到最终的规律。因为它获得的数据是全体的样本数据，从巨大的样本数据中进行分析总结，所以它能够允许不精确，但一定要有足够多的数据量。并且，它也不需要数据的来源（比如用户）具体回答什么问题，而是实打实地去获取用户的“一切行为”，记录他们的全部信息，并一样不差地全部复制过来，变成用以分析的参考数据。

大数据不仅让我们不再期待精确性，也让我们无法实现精确性。当然，数据不可能完全错误，但为了了解大致的发展趋势，我们愿意对精确性做出一些让步。如果说哪天技术变得完美无缺了，不精确的问题也就不复存在了。错误并不是大数据固有的特性，而是一个亟需我们去处理的现实问题，并且有可能长期存在。如今，大数据给我们带来的利益，让我们能够接受不精确的存在了。

假设你要测量一个葡萄园的温度，但是整个葡萄园只有一个温度测量仪，那你就必须确保这个测量仪是精确的而且能够一直工作。如果变成每分钟测量十次甚至百次的话，不仅读数可能出错，连时间先后都可能搞混掉。因此我们为了获得更广泛的数据而牺牲了精确性，也因此看到了很多如若不然无法被关注到的细节。如果每隔一分钟就测量一下温度，我们至少能够保证测量结果是按照时间有序排列的。

试想一下，如果信息在网络中流动，那么一条记录很可能在传输过程中被延迟，甚至干脆在奔涌的信息洪流中彻底迷失，在其到达的时候已经没有意义了。虽然我们得到的信息不再那么准确，但收集到的数量庞大的信息让我们放弃严格精确的选择变得更为划算。

再假设如果每100棵葡萄树就有一个测量仪，有些测试的数据可能会

是错误的，但众多的读数合起来就可以提供一个更加准确的结果。而它提供的价值不仅能抵消掉错误数据造成的影响，还能提供更多的额外价值。因为这里面包含了更多的数据，也不会更加混乱。

凯艾尔说，我们为了高频率而放弃了精确性，结果观察到了一些本可能被错过的变化。虽然如果我们下足够多的工夫，这些错误是可以避免的，但在很多情况下，与致力于避免错误相比，对错误的包容会带给我们更多好处。

有时候，当我们掌握了大量新型数据时，精确性就不那么重要了，我们同样可以掌握事情的发展趋势。这又是一个关注焦点的转变，正如以前，统计学家们总是把他们的兴趣放在提高样本的随机性而不是数量上。因为在进行数据转化的时候，我们是在把它变成另外的事物。

然而，除了一开始会与我们的直觉相矛盾之外，接受数据的不精确和不完美，我们反而能够更好地进行预测，也能够更好地理解这个世界。因为拥有更大数据量所能带来的商业利益远远超过增加一点精确性，所以通常我们不会再花大力气去提升数据的精确性。

大数据的非标准性，迫使我们要讲究效率但可以不追求极致精确。

● 要知道，95%的数据都是非标准化的，5%的数据是标准结构化数据。

● 大数据处理要考虑全部数据就要接受非标准数据，不能以部分代替全局，数据分析的一个必经过程就是将混杂的非标准化数据标准格式化。

● 网络上的贴标签方式就是很好的归集到标准化数据上的一个例子。因此人们需要收集纷繁复杂的数据。

☆描述性的分析

什么是描述性的分析呢？

通俗来说，就是我们常看到的报表、图标、统计图等等。我们期望通过描述性分析来了解过去发生了什么，为什么发生，以及了解现在正在发生什么乃至未来会发生什么。然后进行理性的思考，我要做什么样的事情，我想要未来发生什么，能够在未来让这件事情发生。

也就是说，在最好的情况下，我们能够将描述性分析对未来做出某种预测，并且保证预测的精确性。

☆实时性

对于任意数据来说，实时性都是非常重要的。

它不仅仅是一大类的思维和方法学，而且实时性一定比绝对的精确性更重要。众所周知的购物篮分析，就是基于历史的数据做出相对精确的分析。最好的时机是用户还在浏览、找东西的时候，而不是最后结账

的时候，所以这是当你在超市购物的时候所能想到的一个非常实用的问题。

CHAPTER 7 大数据与生活变革

云分享与云消费

大数据时代给我们的生活带来的好处当然是显而易见的。现在，我们人人拿着一部手机，有的人甚至好几部智能手机；我们的面前也摆着电脑，并随时可以上网；我们面对爆炸式的信息，遨游在信息之海，可轻松地获取数据，来改善生活的质量，享受科技带来的乐趣。

数据爆炸引发了生活变革。这不仅使我们的世界充斥着比以往更多的信息，而且其增长速度也更快，快得让人感觉眼花缭乱，应接不暇。这种信息总量和速度的变化，最终导致了信息形态的变化，从量变引发了质变。

☆云分享——史无前例的信息扩散速度

第一个与生活有关的质变，就是云分享的出现。

通俗的解释是：云分享是依托互联网各个终端的用户作为一个“数据库云”，并且把自己喜欢的各种资源，包括视频、音频、图片、文字等格式的资源，以一种非常统一的方式分享到数据库，形成独立并联结于外界的数据节点。

这时，其他用户如果有需要，就可以通过互联网上的查找工具方便快捷地查找到这些资源，然后进行共享。需要担心安全吗？不需要，因为每个人所提供的分享的资源都是通过某一安全协议来实现的，互联网技术已提供了足够强大和稳定的保护机制。

这一理念出现于2009年，谷歌等巨头都投入到了对于云分享的研发和推动。目前，它在中国的发展也十分蓬勃。比如现在十分流行的“云推送”应用，人们在电脑上看到有趣的信息要想给朋友分享时，只需要通过右键选择这一功能，就可以实时地推送到好友的手机或者邮箱之中，对方什么都不用做，就可以分享到你提供的信息了。

云分享的这个特点让人们感受了自由，因此它的受众极广，对于人们的生活所产生的影响也极为巨大。可以说，云分享是大数据引发的第一场生活变革，它真正地让我们的生活、兴趣与别人的需求连接到了一起，构成了一个共同的世界，推倒了之前那堵厚厚的封闭的墙。

☆云消费——每个人都是一个消费信息终端

云消费是什么呢？

它首先体现在消费的实时性和无阻碍上，是指利用云计算、电子商务等技术，通过云整合团购网站、运用云物流等手段，能使供应链的团购端向外来延伸和发展，建立起基于电子商务云的集中的团购平台，从而达到构建安全、稳健的供应链渠道，既服务消费者，又实现超大规模

销售的商业目的。

比如，当我们每一个人都拥有了一部智能移动终端+移动互联网+SNS账号的时候，人们通过购物搜索和消费体验所分享的动作，就让自己成为了一个真正的“消费信息终端”。在这时，人们既可以通过SNS获得其他人的消费建议，也可以向他人分享自己的消费体验。

“云消费”时代的典型特征：

全裸时代

就像扒光衣服的圣徒一样，企业在市场面前、在消费者面前没有秘密可言，任何一方面都处于（经常是被迫）公众的注视之下。这包括产品的生产过程、供货的流程、成本甚至企业高管的道德品行。

在全裸时代，有什么东西可以逃过公众的监督吗？没有！云消费和云分享的环境是如此开放，其威力是如此之大，让消费者通过这个环境都拥有了一副透视镜，能够全方位地监控企业。

风暴特征

在“云消费”时代，企业如果发生了危机，即便这种危机不是它造成的，它也会立即陷入一种噩梦般的环境之中，引发大范围的“声讨”。就像一场风暴，它的后果是摧毁性的，会延伸到媒体环境和政治环境，波及大量的消费者，并造成极大的压力。

风暴特征还具有四个特点：一是参与度高，社会成员人人参与；二是传播面广，业界、媒体、政府、公众，传播到每一个角落；三是影响大，可迅速成为社会热点，人人关注；四是破坏力强，就像一场无可抵抗的风暴，企业一旦控制不好，后果极其严重。

国内有一些很典型的例子，比如三鹿奶粉事件就具有这种风暴特征。

骨牌效应

骨牌效应讲出了云分享之于云消费的辅助效能，当危机发生时，由于分享的速度极快，引发人肉搜索或集体抨击：人人来推墙，就像推倒了骨牌，导致某个行业的全面危机，波及无辜不说，还会对大众的生活造成深远的影响。

这三个特征首先给企业的品牌管理带来了巨大的挑战。第一，品牌的内涵有时不太重要了，或者说在某一些特定的时期将不被人重视，而是被口碑替代。人们说你好，你就好，说你不好，你无力辩驳。第二，品牌管理的操控性被极大降低，在对品牌形象的塑造中，只能诚实对待，说谎的风险被无限扩大，因为人们通过大数据的技术手段，很容易就能发现真相。第三，企业在进行品牌管理时，要采取更加聪明的做法，既不能无所作为，又不能过于虚伪和夸张。

这其中既有机会，也有挑战。但是有一个方向是肯定的，那就是：企业的核心价值观将会越来越重要。你怎么说并不重要，重要的是你怎么做。好的企业一定拥有自己的核心价值观，非此不能在激烈的竞争中生存。比如美国的苹果公司，它是一个高高在上的榜样，令全世界几乎所有的企业可望而不可即。

我们的“私人订制”

我们在看视频网站的时候，通常会在页面的底栏或者侧栏看到有个“猜你喜欢”的选项，你仔细一看发现：咦？我还真的喜欢。这些网站似乎有知晓我们喜好的能力，而且每个人都是“私人订制”，比如你喜欢综艺节目，“猜你喜欢”里就有综艺节目；你喜欢恐怖电影，“猜你喜欢”里全是恐怖电影。

大数据进入了“人性化”时代，更是一个“私人订制”的时代。不仅是“猜你喜欢”，而是“我知道你喜欢”。

不过，关于大数据，对于更多的普通人来说，仅仅是知道这个抽象的概念，至于大数据究竟是怎么回事，大多数人的感觉更像是“听说过，不大清楚”。

大数据发展到现在，作为普通用户也可以利用大数据的入口。在2013年底的时候，视频网站爱奇艺率先让用户“尝了鲜”，如果你最近用爱奇艺关注了综艺节目，就会发现在当前播放节目的进度条下有两个虽小但很实用的小字：绿镜。点击之后，你就可以看到剪辑过的节目片段合集了，这就像一部电视剧的精华版，绿镜替你删减了你可能不喜欢的情节。就拿时下最热门的综艺节目《爸爸去哪儿》来说，这个节目一期的总时长是90分钟，开启绿镜之后，你会看到一个仅有29分钟的精华版。

这个精华版，是由网站的视频编辑剪辑的吗？

当然不是，是后台系统进行的，也可以说是所有观看视频的普通用户一起做的。因为我们在看视频的时候，通常都会根据自己的兴趣点击暂停、快进或者倒退键，不要小看了这些动作，视频内容的好坏正是通过这些无意识的点击动作体现出来的。

“用鼠标来反映哪些是他们认为好看的、哪些是无聊的。”爱奇艺的首席技术官汤兴说。

而绿镜呢，就是所有用户“评价”的汇总，后台系统经过一系列运算，把不喜欢的过滤，把最受欢迎的剪辑出来，从而形成符合绝大多数人口味的精华版。

我们刚开始提到的“猜你喜欢”是传统的算法，绿镜在这方面做了创新，比传统方式更简单更实用，第一天的用户点击量就超过了20万，而且是在没有经过宣传的前提下。我们可以预测，很快就会有更多这样的大数据产品出现，这不仅节省了用户的时间，而且给节目制作方提供了有力的依据，他们将更加精准地把握人们最喜爱的内容。

私人订制的时代来了，而且是悄悄来临的。

绿镜的诞生其实是源于“吐槽老板”，刚开始不过是几个工程师开的玩笑。2013年2月份的时候，湖南卫视的节目《天天向上》邀请了爱奇艺的CEO龚宇作为嘉宾，与此同时还有电视剧《笑傲江湖》的两个主演霍建华和陈乔恩。第二天，这个节目在爱奇艺上线，几个工程师出于某种好奇心偷偷地在网站后台运用算法做出了这一期节目的用户行为数据。

结果，如大家猜测，当他们的老总出现的时候，节目收视曲线狂降到谷底，而到了两位《笑傲江湖》的主演出场后，曲线陡然升上高峰。收视曲线反映了一个很“残酷”的问题，一个嘉宾究竟受不受欢迎，不需要做民意调查，看看曲线一目了然。

就是这次意外的观察成为了绿镜开发的灵感来源。爱奇异的工程师花费了两个多月的时间，终于做出了绿镜的算法。

通常，一个产品的研发需要几个部门的分工协作。就以《爸爸去哪儿》为例，如果要产生一个绿镜版，首先需要收集数据，然后进行清洗，最后根据清洗后的数据建立模型，完成视频的编辑。这里的清洗就像过筛子一样，是滤掉那些“噪音”数据，这些数据并不是真实的用户行为。比如一个用户在某一点上暂停几秒钟，他可能是在截屏，这种数据就是无意义的；而如果他在一个点上停留了几分钟，那他可能是去上厕所了。

据绿镜的研发者说，他们在建模的时候，对于快进、快退、暂停、分享、评论以及截图等行为会有不同的权重，然后根据重要度为每个片段打分，这个过程是最为复杂的，而之后的计算过程就非常简单了。比如一个30分钟的视频，只需要十几毫秒就能做完。

相对于传统行业，互联网对用户数据的收集要容易得多，因为用户的所有交互行为都会在互联网上留下痕迹，这些数据并没有被抹去，而是存储在各家公司庞大的数据库里。就像绿镜，需要生成绿镜版本的前提是一个节目要有10万以上的播放量，这个数量很容易达到。

通常，那些一线综艺节目整体播放量的单位是亿，比如《爸爸去哪儿》，由于热度极高，基本上上线一小时就可以达到生成绿镜版本的数量要求了。而且，每隔20分钟，后台系统还会重新更新一次。

由于大数据刚刚兴起，绿镜也尚不成熟，所以绿镜版本的节目有一个很明显的问题——不够连贯。也就是说，虽然视频片段被剪出来了，但是并不能衔接流畅地播放。由于这个局限，绿镜暂时只能用于综艺节目。综艺节目对剧情连贯性没有太高的要求，观众主要看的是自己喜欢的话题或者表演嘉宾，有些片段全部截取掉也不会有影响。而电视剧或者电影则对连贯性要求很高。

当然了，绿镜的版本仍然在继续优化。由于内容的连贯度并不容易解决，所以研发团队着重把精力放在满足每个用户对时间的需求上。如果你是比较繁忙的人，也许只需要通过绿镜版本了解下大体内容就可以了，而对于那些有大把时间，特别关注某个综艺节目的观众，就需要看完整版。

我们每个人都是大数据世界里的一个小分母，同时，我们也都是有权享有个性化服务的VIP。

据汤兴说，绿镜的下一步将会进入新闻和体育节目领域。新的版本可以让用户自己设定看节目的时间，比如早上起床刷牙舒展身体的时候想要来一次10分钟的要闻浏览，那么用户就可以选10分钟的选项，在这10分钟里可以看到最精彩的新闻。如果你是一个足球爱好者，一定会喜欢这个版本，因为一场90分钟的球赛，真正最精彩的往往就10分钟，大多球迷只想看进球。

其实，对于爱奇艺来说，绿镜还可以衍生出另一个产品——更精准的收视数据，很多制片人对此求之不得。

比如，在北京的郊区有一个影棚，制片人王凡做了一档叫作《大王小王》的情感访谈节目，研究上一期节目的收视率则是他必须要做的功课，王凡的数据就来自第三方机构。

问题是，长久以来的统计方式都无法让业内满意，其衡量收视率的数据准确度并不高。

传统的收视率调查采用抽样调查方式：在一个城市里，根据当地居民的性别比例、年龄分布以及职业和收入情况等，选取一定数量的样本户，通过对样本户收看内容的监测，来推算整个地区的收视情况。

其实样本户的数量非常有限，比如一个城市有上千万的人口，但样本户的数量通常不会超过500户。而如果有样本户受到贿赂，将收看的频道固定在某个电视台上的话，就会影响到整个地区的收视率。一直到现在，虽然并没有实际的样本户受贿的案例被爆出，但电视台之间的互相指责却从来没有停止过。这就从一定程度上说明了很大的问题。所以对于百度等互联网巨头来说，这是一个很大的市场。

绿镜数据现在只是作为增值服务提供给合作方，还没有作为产品单独出售。比如爱奇艺在购买《爸爸去哪儿》等网络综艺节目版权的同时，会把绿镜数据的分析结果作为附赠品送给电视台。

这就是绿镜的优势，取样数量很大，数据分析精确度高，一些热门视频的点击量甚至可以达到上千万甚至上亿。随着绿镜精准度的不断提高，这些视频网站的数据慢慢地就会影响到上游的影视内容制作。

很多视频网站都在做这项内容，他们在购买了节目之后，向制作方

提供整体的播放量、网友反馈以及某一节目类型的偏好关联，根据这些数据说服制作方对节目内容做出调整。当然了，这需要一定的时间，毕竟在传统制作和大数据的运用之间，还存在着很多的问题，大多数的制作方还是更偏向于传统的做法，比如王凡那样的传统制片人，他会更喜欢央视索福瑞的数据，因为视频网站提供的数据对他来说缺乏权威性。

汤兴讲了一个用户脸谱的概念，他说爱奇艺下一步会细分用户群体，涉及的内容包括观众的性别、年龄、所在地等等。因为针对不同性别和年龄段的用户，喜欢的题材存在不小的差距，地域性也很明显，比如北方的观众更喜欢郭德纲，而南方的观众会倾向于喜欢周立波。脸谱划分完成后，不同的观众收到的推荐内容是不一样的，就算是同一个用户，在PC和手机上的推荐内容也不同。脸谱划分可以提供更精准的数据，仅凭这一点就能够吸引大量的广告投放。

在此之前，各大门户网站在引导用户看节目的时候，通常是后台编辑根据自己的喜好来推荐，在节目上线之前，他们需要看上数遍，从中挑选一些关键点，然后做出吸引人的标题标注在进度条上。“猜你喜欢”就是这个原理，只不过参照的数据是同一个用户在一段时间内的动作。比如一个用户在看完这段视频之后接下去搜索到的视频，数据上关联度越高，就越容易被列到推荐列表上。

大数据的算法与点击量息息相关，每一个算法工程师可能只是将推荐的成功率提高零点几个点，但对视频网站整体的流量贡献却可能是几百万、几千万。

透明社会——隐私大爆炸

大数据在带给我们便利的同时，有一个很大的隐患问题——隐私大爆炸。举例来说，假如你删除了电脑上的某个文件，清理了浏览痕迹，觉得自己的隐私得到了保护，但事实上，你的每一个互联网动作都被存放在后台的一个数据库里。世界上的各个角落都有无数台摄像机进行二十四小时的公众区域监视。

这些巨大的商业数据库能够探究你的财务状况，而且只要他们愿意，可以把信息卖给任何愿意付费的人。在全球互通的信息网络里，你的每一次浏览都被记录下来。想知道你的车子开到哪里去了？智能型收费道路就能做到。

日新月异的新科技在更加严重地搜刮我们的隐私。

活在这样的一个世界，你会感到紧张吗？很难想象人们的答案是否定的。

在大数据时代，我们每个人都是完全透明的人，没有任何隐私可言。只要掌握相关技术的人想这么干，他就可以把你的一切都晾晒在阳光下面，让全世界的所有人看到。

海量的数据，正让监控变得轻而易举，而且监控的成本越来越低。

特别是如今我们生活在一个数字化的世界，每个人都像一张名片，被标注上符号，存进数据库。试想一下，你常用的身份证号、银行密码、各种通讯工具的账号密码、手机号等，这些号码如果被一些你并不认识甚至想象不到的人利用，那么，你还有隐私吗？

据悉，我们这个时代的数据量正在以指数形式增长。去年这个数字已经达到了2.8ZB(1ZB=1024GB)，而到2015年，这个数字还会翻一番。为这些数字做出贡献的，有3/4的比例是个人创造和数字文件的移动。比如，一个标准的美国上班族每年可以贡献180万MB的数据量，平均每天有约5000MB，这其中包括下载的电影、文档、电邮以及这些数据通过移动或非移动互联网传播时所产生的附加数据量。

这些数据看起来像一盘散沙，不具备任何关联性，然而在庞大的运算能力面前，这些海量数据就得到了高效的整合，它们从每一个小小的细节里体现出了联系。

“棱镜门”事件的爆发，所折射出来的就是对个人隐私权漠视的最强大抵抗。

现在越来越多的人喜欢用Facebook，然而你知道吗？Facebook已经实现了对个人信息收集的自动化与实时化。据首次公开募股时的财务档案显示，Facebook上每位用户的图片和视频资料数据量约为111MB，而

Facebook的用户数如今已经超过了10亿，算下来，这是整整100PB(1PB=1024GB)的个人信息数据。有了这些数据，我们甚至可以窥探出一个人的未来。

来自美国罗彻斯特大学的亚当·萨迪克和来自微软实验室的工程师约翰·克拉姆发现，他们最多可以预测到一个人在未来80周后可能到达的位置，准确度高达80%。为此，他们在32000天里，收集了307个人和396辆车的GPS数据，并以此建造起一个“大规模数据集”。

在西方国家，消费者的信息监控已经发展为一项规模达几十亿美元的产业。其中的企业基本不受什么监管，即使你是一个很有影响力的人，你的个人信息在交易的过程中卖价也不会超过一美元。

监督和被监督的力量不平衡，那些手中掌握着更强大数据分析能力的大公司以及更强大的政府，因此就拥有了自由利用这些信息而不受监督的能力。

“棱镜门”事件所折射出来的，就是这一潜在的危险。所以不得不说，伴随大数据时代的到来，我们面临的挑战就是需要建立一套新的监督制衡机制，这样才能规范政府的行为，从而建立一个更加开放的社会，减少大数据错误带来的危害。

生活在大数据时代，人们的隐私不经意间就会泄露，传统的保密方式失去了作用。这就要求那些保存和管理信息的企业需要拿出承担责任的态度和勇气，他们将面临更大的责任，而这也应该成为一种新的隐私保护模式：政府不应假定消费者在使用企业的通讯工具等产品时主动透露了自己的隐私，那样就意味着他们授权企业使用这些隐私。

更大的力量意味着更大的责任，现在已经到了那些掌控大数据的大企业和政府负起责任、构建安全、完善保护网的时候了。

面对数据化生活，你做好准备了吗？

20年前的世界，古人无法想象。而今天的世界也已远不是20年前的世界。现在，大数据将掌握我们的一举一动，甚至能够预测出我们的下一举和下一动，这一点也不夸张。

所以，一个很严肃的问题就是：“数据化生活，你真的准备好了吗？”

在大数据时代，我们面临着太多的选择。比如早晨起来，该喝什么牌子的牛奶，看什么样的新闻，听什么音乐，甚至于用笔记本还是台式电脑工作，都会构成一种选择的苦恼。与此同时，由于信息太多，更新太快，我们也面临着太多的陷阱，以及海量的商业信息轰炸和各种身份的人向你发送信息。这些不但改变着我们的生活，也在考验你的独立思考、理性决策和创造的能力。

你将如何应对这些目不暇接的变化呢？

我的建议就是，即便你不确定自己将如何应对，但至少不要逃避。当生活正大踏步地步入数据化时代时，它也为我们提供了无限的机会，让我们能够利用大数据技术，改善和提升生活的质量。你既可以针对自己，也可以去帮助别人。

比如，牛津大学就成立了英国首个综合运用大数据技术的医药卫生科研中心，通过搜集、存储和分析大量医疗信息，确定新药物的研发方向，从而减少药物开发成本，同时为发现新的治疗手段提供线索。

还有，美国普林斯顿大学的一些学生针对独居老人展开了一项“魔力地毯”计划。具体做法是，在普通的地毯上安装可以记录老人脚步的传感器，当老人在上面走过时，相应的数据会被上传，电脑收到并分析这些数据，然后与老人健康时的脚步相比较，以此来分辨是否需要就医。如果需要，就会自动触发警报，把信息实时发送给附近的医院。

这些都是伟大的创举，这表明除了为企业创造利润，大数据也能无孔不入地改善我们的生活质量，为普通人提供便利。要知道，无处不在的海量信息现在正改变着整个世界和我们的日常生活方式，一场难以言喻的大数据革命已经悄然来临了，它正在接管与掌控我们的生活。

你必须做好准备，而且你也应该积极地行动起来，让数据化生活成为自己的新习惯，改善自己的生活、工作和社交圈。

CHAPTER 8 大数据与社交变革

颠覆性的社交理念

在社交领域内，我们能想到的第一个概念就是“关系”。关系并不局限于我们所认识的人，比如朋友、亲戚、同事和客户。这些直接关系的“关系”，也是我们的人脉资源，它可以启动一种链式反应，并最终连接到我们，就像是遥远的波浪在力的作用下向我们慢慢涌来，然后逐渐到达我们的身边一样。

传统的社交理念是碎片式的，就是我们只跟直接关系有联络，然后再通过他们去认识他们的人脉资源，就像一片片的叶子，通过互相之间的枝脉相连，建立一种间接联系。比如六度理论所言：“通过六个中间人，你可以联系到世界上任何一个人。”我们与最后的目标，中间还有六个分隔，这就是碎片式的。

六度理论在纸面上十分理想，但是在现实当中，可以和任何人建立连接的梦想恐怕很难顺利地实现。

原因在哪儿呢？在于每个个体的影响力不足，虽然理论上是六度连接，但个人的辐射力只有三度，甚至只有两度。超过了两度或者三度，我们的影响力就会失去效力了，难以与第四度以上的人建立亲密的直接的连接。

因此，结果就是，我们认识相距三度以内的人的概率大大高于相距三度以外的人，我们请求得到的帮助也往往仅限于三度之内。

大数据时代改变了这一传统社交理念，将碎片式的社交连接变成了网式关系库。什么是网式关系库呢？就是“点对点”的直接连接，我们不需要再通过这六个人，而是在大数据工具的帮助下，直接与目标关系人建立联系。

最近10年来，各种新的信息传播工具随着技术的发展在中国迅速涌现。智能手机和3G网络就不说了，它们不但催生了大量的新兴传播媒体，也掀起了一场社交领域的革命，并由此颠覆了旧的社交理念。

比如微博和微信，就是这样的大数据工具。

☆“双微”操纵社交——微博和微信

微博既是一种社交工具，又是一个媒体平台，这使其具备了丰富的多面特质。当然，微博也是一个信息平台，这一点毫无疑问，人们也更多地把它作为一个信息发布平台来使用，至少人们在开始使用微博时的第一印象是这样的。

但在我看来，微博的信息平台背后其实隐藏着媒体和社交网络的性质，尤其后者，功能十分强大。

我们日常可以接收的信息有两种。第一种信息是一般化的，就是我们只关注信息本身，不需要在意这个信息的来源是谁，比如新闻，我们看到了一则新闻，只关心新闻的内容，而不太关注是哪家媒体发布的；第二种信息是特定的，我们既关注信息本身，又十分关注信息源，即由谁发出的，比如对于某个事件的评论，我们既了解评论，又会关注这个评论由谁发布。

人们在接收特定信息时总是会十分在意信息的来源。来源不同，意义往往不同。

比如一个普通人说明天可能会地震，你不会相信，但如果是某一个地震专家所说的，你可能就会深信不疑。

同样的一句话，不同的人说出来，意义完全不同。而对于不同的人，也会有不同的意义，就像你讲了一句话，有的人可能当成垃圾信息，但有的人可能如获至宝。如果信息源不明的话，我觉得相当多的信息对我们就没有什么意义。

发布一般化信息的，我们就可以称之为“媒体”，微博就有媒体的功能；而发布我们十分关注的特定信息的，就可以称之为“社交网络”或者社交对象。对后者的微博账号，我们就当成了一种社交资源来使用了。虽然其中的区分并不严格，但大体已经向人们展现了这两种不同的性质。

这也意味着，微博更多是一种大众化的社交工具，因为大家都想认证，以使自己的微博账号更加有名，获得更多的关注。以这一目的为性质的移动社交正在颠覆人们的生活，在最近两年间，逐渐达到了一种顶峰。

微信呢？与微博相比，微信虽然也有媒体作用，但它的社交功能更加明显，所以发展势头迅猛，一枝独秀。从发布的那天起，它就表现出了一种社交垄断的迹象，比如腾讯公司发布的数据显示：从0到1亿用户，微信用了14个月时间；从1亿到2亿用户，用了6个月时间；从2亿到3亿用户，只用了4个月时间。然后，微信的用户数量稳步地以每5个月增长1亿的速度迈进。

到2013年的10月份，微信的用户数量已经突破了惊人的6亿大关。单纯从数字上看，这意味着有一半的中国人都在使用微信（当然其中包括许多国外用户，甚至在叙利亚内战中，无论是政府军士兵，还是反对派武装分子，许多人都在战场上使用微信进行联络，这真是一种惊人的现象）。

另外，腾讯公司的大社交战略也犀利地扩张到了商业应用中，微信是它当仁不让的前锋将军，是开拓市场的主力。比如微信的5.0版本，

除了公众号折叠、扫一扫等常用功能的改善以外，还加入了重磅的游戏中心、支付功能等新的涉及商业的应用。

这样一来，就彻底地改变了微信的起初模式，在社交功能的基础上，使其摇身一变成为全新的商业、生活、娱乐综合体。其涉及的范围从购物、政府公示、生活类服务到保险理赔、金融理财等领域，几乎无所不包，成为了人们的现实生活在移动互联网上的全面延伸，让人足不出户就能用微信解决一切生活需要。

而且，在微信提供的所有服务中，它是唯一的入口，是排他性的。使用这种手段，微信对自己涉及的市场形成了垄断，产生了闭环效应。它展示出了其在移动端最大的统治力，将线下零售商、线上电商、家电、视频等企业通过微信的“语音识别”、“地理位置”、“上传图片和视频”等九大接口应用连接起来，让人看到了一个真实可触的“O2O”商业帝国。它的目标是打通线上线下，形成移动端的生态系统，随时创造并且满足用户瞬间产生的需求。

最重要的是，用户始终会注意到，微信是其中唯一的入口。这是更多作为媒体与社交平台的微博做不到的。

☆“来往”——阿里巴巴的大社交工具

在大数据的社交理念的推动下，马云也有所行动。2013年9月，阿里巴巴推出了“来往”，希望以此打破微信的垄断地位。为了在短时间内达到较高的增长速度，阿里巴巴和马云投入了相当大的精力，他们主动扮演了一种挑战者的角色。

马云说：“我们要挑战微信，要把不可能变为可能。”整个阿里集团都开始将移动战略的布局重点放到“来往”的身上，因为他们早就感受到了微信的巨大压力。

所以，“来往”推出两个月后，注册用户数就突破了1000万，日活跃用户数增长了500%。“来往”用户建的扎堆数也超过了10万大关，其中千人以上的大扎堆就超过了1500个。这是在社交大数据发展的大背景下，马云所能做出的必然选择。阿里巴巴必须加入这场竞争，否则就会在失去先机的情况下，丢掉未来的追赶机会。当然，这也会让我们的未来更加精彩，因为微博、微信和“来往”无疑是近几年来大社交理念逐渐成型的一个缩影，它们之间的竞争越激烈，人们能够享受到的便利也就越多，对社交的好处也就会越多。

☆图谱分析——从被冷落到大数据热点

社交工具的集合，有一个特征是必须引起重视的，就是对于关系图谱的分析。

在以前这是一个备受冷落的领域，但随着大数据的兴起，已变成了

大数据分析的热门领域，主要被用来分析数据节点之间的关系和相似度。体现在社交领域，就是社交网络的人际网络关系图谱分析，虽然广义上的图谱分析的应用范围要更为宽广。

它的一个最广泛的应用，就是通过社交技术来串联和展现的营销价值。至于是否真的能够做到精准营销，以及如何把握机会，把社交市场成功地升级为营销市场，体现社交工具的营销价值，还要看具体的操作手段。

相关性的力量

进行“实时搜索”是当前的一大热点，这实时展示了人们在社交网络中所从事的一系列活动信息，它通常被称为“活动流”，是互联网公司重点收集的数据。实时搜索曾一度被称作社交搜索，但如今它正在逐渐发展壮大，被应用于几乎所有的以网络为应用平台的领域。

于是，社交的相关性排序就产生了，这时当你开始搜索“活动流”时，你所得的结果呈现在页面上，就不会再按照时间顺序排列，而是根据新的需求，或者说形式——每条信息与用户的社交图谱之间的相关性——来排列了。

也就是说，那些与你关系更为紧密的人将会排在前面，那些对你来说更重要的事情也将排在最前面，总是让你优先看到，然后优先处理。

☆以我们的好友为基准

我们将自己所关注的人排在搜索结果的前端，既符合习惯，也是一种人们普遍愿意采用的显而易见的做法。国内的社交平台大多推出了此类应用，但国外的推特还没有采用。所以，当你在推特上进行搜索时，你所得到的结果仍然会按照时间的先后排列。

推特是如此的“固执”，以至于背离了大数据的相关性原则。因为搜索结果中的多数信息都来自于陌生人而不是熟人。我们都知道，如果所列信息来自于自己所关注的人，那么搜索结果就将更为有用。与此同时，我们也愿意在第一时间知道好友最新的想法，看到他们最新的作品或者观点。

另一家社交平台FriendFeed的做法是采纳了这一模式，它尊重相关性，也愿意让人们的社交工具与自己好友的联系更为紧密。比如，FriendFeed会根据用户的社交图谱（联系人和亲密程度）对搜索结果进行特定的过滤，按用户的意愿进行排列。从技术上这并不困难，也符合用户的心理。

人们体验到了相关性的好处。一方面，我们了解到了自己关注的人；另一方面，相关性的数据搜索让信息的沟通渠道更方便了，还有利于整合我们的社交图谱。因为第一时间了解好友的信息，可以帮助我们分析他们的状态、想法和对未来的打算。

一个更实用的好处是，我们可以凭此判断：“现在我和他的关系怎样呢？他是否正在冷淡我？”

这种方式听起来很棒，益处也令人向往，但存在一个问题。比如当你搜索某个人时或许效果很好，因为你的这位好友最近刚出了一本新书，还在书中提到了你与他的友谊。你轻松地通过相关性发现了这一

点，但是很多其他的关键词却无法返回任何结果。原因很简单，你在这个平台上的好友以及你所关注的每一个人不可能对你所感兴趣的每个话题都过来发表评论，你就很难据此找到更多的可以依赖或进行辅证的观点。

☆为了寻找更多的信息源？

很明显，我们的目的不是欣赏相关性的神奇，而是获得更多的信息。这一解决方案也十分容易，就是整合其他的可以信赖的资源，比如拓宽我们的社交图谱。说白了，我们要不断地扩充联系人，让信息来源更多，让自己的人脉资源库更多元。这样你就能接触到不同的观点，去与很多自己缺乏了解的领域进行碰撞，汲取新的思想。

比如，你所搜索到的结果列出来的内容未必来自于你直接关注的人，但它会包含那些你关注的人所关注的其他人。相关性在这时得到了更为直接的体现，也就是说，你的那些“好友的好友”会提供更多的信息，这里面有你需要的部分。

虽然你自己并不熟悉这些人，也并不信任他们，但是他们提供的内容所具备的价值是相同的，并不因为你的怀疑而失去其积极的意义。相关性的意义就在于，我们能看到这些人的观点，然后为继续了解他们而打开一个窗口。接下来的选择就看你自己的了，你可以继续走近，当然也可以保持距离。

还有一种方法，就是建一个圈子，在平台中把与自己兴趣类似的人整合到一起。所有的社交工具都有这种功能，比如微信和微博的朋友圈、腾讯的QQ群等，都以不同的形式满足了这一功能。这种方法已经变得十分普遍，能够帮助你持续地了解除了自己的好友之外还有哪些人是与你类似的。

当然，这么做的成本是比较高的，耗时也长。不过，随着时间的推移和技术的进步，这一难题必然会得到更好的也更为根本的解决。

沟通数据化

未来，我们的沟通也会数据化吗？

没错，在大数据时代悄悄地降临我们的生活的时候，每个人都与之密切相连。很多人更加喜欢通过社交网络工具与他人交流，表达自己的喜怒哀乐，你经常使用的一个表情或者在一段时期内最频繁提及的一句话，你的网购喜好，是喜欢买图书、电子产品还是化妆品、衣服……这些看似不起眼的点滴信息，都以数字的形式被记录下来，逐渐地就会形成一个巨大的数据库。

大数据就是通过这种搜集、记录的方式对信息加以分析，从而了解一个用户的喜好，这样就等于是在消费者与产品之间架起了一条沟通的桥梁，产品的设计者了解到这些信息，就能够更好地从消费者的体验出发，开发出为消费者喜爱的产品。在这样的一种背景下，传统的商业规则正在被悄然改写，一种崭新的模式将要出现。

我们以游戏公司为例：传统的游戏公司在生产出一款游戏之后，经过测试发布出去，静静地等待玩家用户的反馈，在一段时间之后，他们会根据这款游戏的火爆程度来决定下一步的推出计划，比如发布同款游戏的新版本或者重新开发一个新的游戏。这种模式具有相对迟滞性和被动性，因为设计师只能凭借灵感来设计一款游戏，如果绝大数玩家和设计师的口味相同，那么游戏就会大受欢迎，而如果玩家不买账，这款游戏就会失败，很快就会在更新快速的游戏市场中被淹没。

那么我们就很容易了解，以往的游戏输出方式太过单一，仅仅是游戏厂家靠灵感和运气单向对用户输出，可想而知，这种游戏火起来的概率能有多大。这就如同你要请一个陌生人吃饭，不知道他喜欢吃什么或者不喜欢吃什么，结果你可能点了一桌子他忌口的菜肴，不管菜做得再好、价钱再高，客人可能一筷子都不动。这有什么用呢？

大数据时代下，这种情况就得到了改观。游戏公司可以从搜集到的数据库里分析出玩家的喜好，而且能够知晓游戏中玩家比较难通关的地方，有些关卡设计得可能不太受玩家欢迎，那么设计者就可以及时地去修改这些不足之处。另一方面，通过对数据的分析，设计者也可以针对同一款游戏同时做出几个不同的版本，以此适应不同玩家的胃口。如此一来，一款游戏的成功率就会大大提高。

大数据也为近年来新兴的网络购物提供了无形的巨大财富。

我们最熟知的天猫商城和京东商城，它们在网购领域都取得了令人惊讶的成绩。为什么能够如此成功呢？也许你应该了解到，我们每次交易都会有交易记录，而这些海量交易留下的数据记录就是对财富增长的

一种最客观分析，经过大数据的分析处理，商家就会了解到人们的购物习惯，从而在推送广告的时候有针对性地选择用户。

有人戏言“天猫和京东比男朋友更懂得女人的size”，其实正是通过大数据处理，分析出了女人们最喜爱的品牌，经常买的大小型号，也得出了消费者更青睐什么价位的产品，男人最喜欢买什么礼物送给女人。这些数据就成为生产商们设计生产之前的可靠依据。

比如，手机生产厂家可以分析出人们对价位的接受程度，究竟是2000元的手机更受欢迎还是3000元的手机销量更好；也可以分析出人们对手机电池容量的关心度，电池容量是否会影响手机销售量；针对女性手机，就会特别关注像素分辨率的高低等等。如果能够准确地把握这些信息，那么做出的产品就很容易成为爆款。

2013年火爆的喜剧电影《泰囧》的成功或许也可以给我们更多的启示。现如今，随着人们电影品味的逐渐升级，很多人都觉得“这年头烂片太多”，更有多数人叹息像《泰囧》这样能够迎合观众胃口的电影实在太少。既然身处大数据时代，那我们是否能通过大数据分析来提高一部电影的成功率呢?

有一家叫作“The-Number”的网站，曾经收集了过去几十年里有关美国所有商业电影的信息，其中包括了各种影片的类型、出演主角、影片预算、票房以及获得的奖项。这么做的目的是什么？

通过数据分析，他们就可以用一种特定的计算方法预测电影票房。不仅如此，对于一部正在拍摄的电影，这些海量数据也能够提供建设性的意见参考。比如，一部电影的某角色如果启用某某演员，票房就会增长很多。这就使一部影片在推向市场之前就充满了主动性，我们甚至可以说，对某某演员启用的过程中，已经在收获票房了。

因此，大数据时代下，数据给我们的决策增加了更多依据，对于未来的电影市场，也会大大地减少烂片的出现。

作为一种数字媒介，大数据正成为广大生产者越来越关注的资源。通过大数据来删减生产环节中可以避免的风险，发现消费者的爱好，改进产品的缺陷，从源头上改善了“闭门造车”的局面。大数据正作为一座沟通桥梁连接在生产者与消费者之间，产品设计者更加从容，而消费者也容易满足自己的喜好。

社交大数据——新的营销革命

这是一场席卷全球的智能广告革命，基于社交大数据的天然优势，无孔不入的广告到处都是。只要我们是“社会人”，还在产生和使用数据，还在浏览网页，使用手机，智能系统就能将各式各样的广告在我们眼前播映。

高效广告不再出自4A广告公司的大牌创意师之手,而是出自由社交数据支撑的智能软件。它自动工作，不需人力；它可以自动跟踪，并将用户近期关注的领域收集并进行分析，然后推送与之相符的广告信息。

脸书打造了一个能取代传统广告代理公司的高精准广告系统，即广告客户只需将数万张产品照片上传至数据库，那么一旦相关的用户登录了脸书，系统就会根据该用户的兴趣特点，自动生成相关的广告，投放的依据是对于用户的社交关系图谱的严谨的数据分析。由于这一技术的发明，在2012年，脸书公司的广告收入攀升到了近50亿美元。

这场营销革命主要体现在三个方面：

☆个性化的用户体验

用户体验本身就是消费者十分追求和在意的东西，如果再加上个性化的特性，它当然会更好地促进销售的成功。

比如，谷歌公司曾经公布了谷歌眼镜的一段新视频，它们对其使用过程中如何进行互动和其他服务功能进行了展示。这副眼镜可以进行移动通讯、摄影、GPS定位，具有十分强大的多用途的功能。虽然对它实用性的质疑之声始终不绝于耳，但谷歌仍然十分坚持，因为它的理由正是基于个性化的用户体验。

在人们对于自己使用或期望使用的产品、系统或者服务进行体验时，总是希望以自己为中心，与自己的个性进行融合。这也是为什么好的公司都开始强调以客户为中心，而不是以产品为中心。

比如，360公司的董事长周鸿祎就不止一次地强调，所有的员工都要像“小白”一样思考，像专家一样行动。为什么呢？因为他希望自己的员工能将每一个潜在用户设想为电脑白痴，创造出最简便、易操作的客户体验。

满足用户的个性需求，这是大社交时代新的营销领域最为关键的理念，也是社交大数据的必然结果。谁的用户体验更好，相应的品牌亲切感就会越高，给用户的印象当然也就越好。

国内还有一个很现实的例子，苏宁易购对于网购业务一直是雄心勃勃，充满了野心，也投入巨大，但为什么就是没有办法战胜京东和当当呢？不是缺钱，而是输在用户体验上。用户在网购时最在乎的体验和最

需要的个性服务是什么？是速度。京东可以限时达，经常在当天就把货送到，苏宁易购却仍然保持着蜗牛一般的物流速度。

对用户来说，在如此快节奏的社会生活中，这样的体验是非常不好的。所以，人们喜欢京东和当当，而不是苏宁易购。

只有当越来越多的中国企业开始真正地在乎用户的个性体验时，他们才能正式地进入大数据时代，才能懂得如何利用社交大数据，针对不同用户的特色需求，提供个性化服务。

在这样一个时代，用户的行为和思维都发生了巨大变化，普通人什么都不做，就已经能够越来越多地影响企业的战略，影响企业现有的业务和未来的发展。毫无疑问，随着社交大数据的发展，用户和商家之间的关系已经发生了革命性的逆转。

在大数据时代，根据社交数据来为用户体验提供更好的服务，就现阶段而言，我们至少在两方面能够做得很好。

第一，以云计算为基础，最大限度地获取整体数据。使用这些数据，来为判断客户的需求与喜好提供参考，进而有助于用户体验的设计。

第二，以社交工具为基础，可以针对用户的不同需求，进行量体裁衣，进行个性化的营销。面对用户需求逐渐细分的市场，凭借大数据强大的数据分析，我们有足够的条件帮助企业对不同的消费者提供个性化服务和营销。

总的来说，个性化的用户体验，就是针对适当的客户，在适当的时候，说最恰当的话，做出最正确的营销。我们要让个性化营销变成用户的亲密伴侣，可以给他们带来快捷、舒适和亲密无间的体验。假如你是一家企业的管理者，你要让自己成为用户的闺密，知道他们在想什么，也知道他们想要什么；不但要了解他们的过去和现在，还要预测他们的将来。

☆实时化的营销决策

大数据对营销最大的改变，是决策的方式与效率。在传统时代，营销决策的流程是相当麻烦的，也根本做不到实时化。比如，企业以前在做互联网营销时，通常是以网站为目标，再由网站锁定目标人群，然后再锁定企业要传播的人群，再去制定相应的营销决策。但在大数据时代，营销是以内容、关键词匹配广告的模式进行的，而且这种模式就是实时化的，也是自动推送化的，结果还更加精准，并且覆盖到所有的人群。

在实时化、精准化和全覆盖的营销环境中，每个用户都无处可逃，只要他上网浏览网页，就会第一时间获知企业的产品。这一目的的实

现，是以新的数字营销技术为前提的，它可以对互联网的用户群进行跟踪和定向，通过海量数据去计算消费者的偏好与兴趣，转换为以消费群体为目标来投放广告，以达到精确地对每个人进行个性化的广告推送。

这时，决策就成为了一种实时的高效率的行为。我们可以实时地监测或者追踪消费者在互联网上产生的海量行为数据，对此进行聚合、运算和挖掘，然后根据挖掘的结果发现结论并做出正确的营销判断。

☆兴趣图谱与消费行为定向

我们的营销工作，从来没有哪一个时代比今天更需要懂得与用户进行分享、沟通的策略。这是因为，用户的兴趣和消费习惯已越来越明显地影响到了企业的销售和决策。假如你不尊重用户的这两个特点，你在大数据时代的新一轮竞争中一定会迅速败下阵来，而且一旦输给对手，就很难再有卷土重来的机会。

在这其中，社交媒体的作用是什么呢？它可以通过挖掘用户的数据制定针对性的沟通策略。哪些用户的数据可以提供这一价值呢？他们在网上所发表的评论、上传的图片、音乐、视频等等，这其中就蕴含着用户的消费倾向。甚至你浏览过什么网页，都会成为这一判断的依据。

美国通用公司的一名营销经理说，社交媒体的本质就是情感维系的交流互动，这些信息是非常宝贵的数据，它们将在营销中发挥越来越重要的作用。比如用户要购买一款汽车，按照以往的方式，可能会是先看一看纸质广告，再去实体店察看、试驾，最后付钱购买。但是今天，用户已可轻松地跳过这一流程，他们从社交媒体及移动终端中接触到其他顾客的评价反馈，从而改变预先的购买计划。

企业呢？这位经理说："企业就需要通过社交平台追踪这些信息，获知他们的想法，来针对性地设计和改进产品，并制定对应的营销策略。"

正因为如此，通过社交网络定向地投放广告成为了一种必然的选择。这就是智能广告开始大规模出现的原因，腾讯公司的一名广告经理就指出，智能广告在未来的发展趋势将向视频广告、微博广告、无线广告、展示广告四个方向发展。而且在中国的网络广告中，视频贴片广告和富媒体广告将是展示广告持续增长的主要驱动力。这些广告都瞄准了用户的兴趣和消费踪迹。

在社交平台，人们分享和交流得越多，留下的足迹也就越多。在这些数据的支撑下，企业也就越能产生出更多的正确的决策依据。大社交数据的核心就是在社交和营销之间建立一座畅通无阻的桥梁，激发人们这种分享的热情，也让企业获得新的营销动力。

企业与用户之间的交流方式正在发生巨大的改变，越来越多的企业

正在主动拥抱这种变化。在我们去年对中国的市场营销人员的一次电话采访中，近70%的受访者表示，推进他们在广告营销领域运用数据管理平台的动力就来自于挖掘大数据的需求。

出自社交领域的大数据应用正在彻底改变全球的营销行业，社交平台已经不仅仅是一种媒体工具或者人脉工具，也不仅是要让人们记住来自不同圈子的推荐信息、新闻话题、情感故事，更重要的是，它催生了人们潜在的消费需求，从而使企业可以方便地记录下他们的消费曲线。

CHAPTER 9 大数据与管理变革

数据说话——更加理性的决策

现在普遍采用的人力资源管理方式大多存在很多问题，诸如主观性太强、单凭经验感觉去判断、信息不对称不透明等，这种人为管理的方式更趋向考察HR的个人能力，而大数据的出现，则为HR头疼的问题提供了强大的解决助力。

美国的IBM公司是把大数据技术应用到人力资源管理等方面的典范。作为一家跨国公司，它将大数据技术的优势发挥到了极致，其创建的Professional Marketplace数据库包含了IBM员工的技能、薪资以及近期日程安排等等可考信息,并且通过数学运算从中攫取到一种资源配置最优的方式。

显然，这对于人们组建团队的帮助是不可小觑的，尤其是项目经理需要组建一支全能团队的时候，就可以参考Professional Marketplace数据库，找到最适合的员工。就像在一堆车票中挑选你最喜欢的位置那么简单。

☆人才测评

人才可以计算吗？当然可以。

在大数据与云计算时代，几乎没有什么是不可以量化和计算的，人才作为一种资源，自然也可以作为数据和计算的一部分。而且，这种用数据说话的人才测评方式，能够帮助HR更轻松地找到所需的人才。

归根结底，人才的管理是资源管理的一部分，从本质上可以实现量化的、数字化的管理。比如财务预算管理掌管一个企业的经济，制订企业的财务计划；人才管理则是企业对于人才的计划，通过量化的方式，对企业的人才进行评估、盘点，从而挑选出目标人才进行培养、训练、提升，最终完全符合企业的要求。

在一个企业当中，现有的人才状况是否与行业要求相符，差距在哪里，有哪些优势和缺陷，是技术差距还是素质差异等等，这些通过什么分析出来呢？当然是数据。如果没有大量的数据分析，仅凭HR的观察和直觉判断，必然存在着巨大的误差，人才管理的科学性也就不存在了。

尤其是当企业发展到一定的规模之后，企业的人力资源管理更是需要科学数据的支撑，对人才的测评通常涉及很多方面，比如人才的绩效数据、人际关系测评数据、技术素质数据等等，企业如果有一个数据库，就可以通过比对数据得出结论，可以有依据地确认企业目前缺乏哪方面的人才，员工在哪些方面普遍存在欠缺，以及哪些人才真正适合企

业未来发展等等。

大数据时代让人才资源管理摆脱了过去传统人才管理的“糊涂”“模糊”模式，进而进入全方位的数字化管理新时代。

作为中国最大的人才管理与测评解决方案提供商，北森公司在几经变革之后，也跻身到了大数据与云计算领域。北森人才管理平台开始采用一种新模式为客户提供服务，比如SaaS模式，正是利用了云计算技术来搭建平台。

这是一个充满了无限前景和潜力的全新领域，也将给HR带来更加开阔和创新的人才管理思维。

☆绩效考核

我想，现在已经不需要为大数据做什么广告宣传了，各大企业的CEO、CFO们也都非常清楚这一点：大数据是这个时代的发展趋势，而且在不久的将来就会成为扼住企业咽喉的生存关键，更是企业制胜的一大法宝。因为大数据打穿了企业传统管理中横亘在量化和非量化、内部和外部、管理和执行之间的IT鸿沟，为企业科学管理决策提供了有力的依据，能够大大地提高企业的决策质量。

绩效管理作为现代企业不可或缺的管理工具之一，其本质就是通过收集各种绩效数据，然后进行分析，从而为企业的各项管理决策提供支持，并对企业未来进行预测以及流程的改进。目前看来，EPM（能效与生产维护管理）显然将会是企业未来竞争的主要目标之一，这也将引发企业信息系统的一次新变革。

现在大多数的企业对EPM的关注重点仍然还集中在对KPI（关键业绩指标）的监测上，这显然是不够的，决定EPM成效的基础前提首先应该是如何选择合适的、准确的关键战略指标。有少部分大型企业已经意识到了这个问题，他们已经开始通过聚合KPI与海量财务与运营数据的数据仓库应用来提取更加深入的分析结果。比如谷歌、亚马逊、eBay和沃尔玛等企业。

传统意义上的企业数据库就像产品仓库，在存储和查询数据的时候完全采用结构化的方式。那非结构化的数据怎么处理呢？显然，传统企业数据库无法做到。而社交媒体和IT技术的发展则使得大数据分析成为可能，大数据工具能存储和分析传统数据库无法处理的海量非结构化数据。

大数据分析意味着企业将能够把支撑决策的数据来源和类型扩展到过去无法企及的领域：通过搜索引擎、社交媒体、博客、视频等结合结构化的交易数据来更好地理解员工的行为。

数据总量和类型的大幅增长

有了大数据分析的帮助，传统的KPI分析工具将具备分析大量的非结构化数据的功能。例如，Net Promoter Score已经能通过分析Facebook和Twitter等社交应用的信息来评估某个品牌或者产品的用户满意度和忠诚度。这仅仅是一个微小的开始而已，随着知识资本向社交资本过渡，一个企业如果无法评估开放社交媒体的信息，也就无法进行有效的绩效管理。

大数据分析的大众化

在过去，企业需要投放大量的资金在IT和数据分析上，以此来获得大量数据的采集、处理和分析。但在大数据时代，这部分资金的投入将得到很大程度上的“减免”，因为有很多数据处理工具是免费的。

Google Trend。趋势性数据的普用分析工具，任何公司都可以使用Google Trend进行大量趋势性数据的分析，并可应用于市场研究。

Social Mention。这是一款可以让企业追踪到社交媒体中的品牌提及信息的工具，甚至能对用户的评价信息进行倾向性分析，比如对某一品牌的评价，大多数用户是反响积极还是普遍不满意。

TripAdvisor。这是一个很有名的在线旅游点评网站，TripAdvisor能够通过专用工具采集旅游者对酒店、餐馆和景点的评价信息，从而向酒店、餐馆的经营者提供一个可参考的数据，便于经营者分析客户的喜好以及消费趋势，从而做出及时的改进。

以上工具都是一些很好的数据分析例子。除此之外，在大数据分析的其他方面，最大众化也是最典型的例子是Kaggle举办的“汽车索赔预测大赛”。这个大赛会召集一些完全不懂业务的汽车爱好者，他们仅依靠厂商提供的有限的数据（具体车型被符号代替），就完全可以预测出不同汽车品牌产品的事故几率。这时你要问了，准确吗？不得不说，这个大赛预测结果的准确性比厂商自己的预测高出340%。

绩效考核的技术变革

当前，大多数企业仍然延续传统，将企业绩效管理看作是基于数据仓库的一种分析和汇报软件，但是在大数据时代，云计算和SaaS（作为服务的软件）正在逐渐成为企业绩效管理软件的技术发展趋势。如果企业不能及时地认识到这一点，必定会在不远的将来被市场淘汰。另外，Hadoop作为一种能够分布式处理海量数据的框架和内存数据分析，也都是企业绩效管理未来的热点技术。

☆人才选拔

传统的人才招聘方式，通常是企业人力资源部门或者政府机构租用场地，承办招聘会，这种方式虽然直接有效，但是其效率以及对企业资源的耗费，不得不说是巨大的。而在大数据时代，随着人才市场丰盛的

大数据新技术的到来，传统的招聘、选拔人才的方式将逐渐淘汰。

谷歌公司每月收到大概10万份以上的简历，如此庞大的数目，企业如果一一进行约谈面试，所要消耗的成本无疑是巨大的。那么，如何在成千上万的求职简历中筛选出最合适的那些呢？

为此，谷歌公司采用了大数据技术，它会让所有的在职员工各完成一份有300道问题的问卷，然后根据问卷的结果分析建立出一套数学模型。很显然，在职的员工经过培训已经在各方面符合公司的要求，这些员工做出的问卷，是最能代表谷歌人力资源需求的。这套数学模型让谷歌摆脱了凭文凭“敲门”的招聘缺陷，从而从其他方面开始去发掘那些潜在的“最适合”申请者，他们也许在学校并不是成绩优秀的人，但他们可能更具有发展潜力。

2003年，美国作家迈克尔·刘易斯写了一本畅销书《魔球——逆境中制胜的智慧》，在此之后，曾就职奥克兰市运动家棒球队总经理的比利·比恩成了人人追捧的明星。

比恩所在的棒球俱乐部在挑选球员的时候依赖于“球探”，而在2002的时候，比恩决定改变这个方法，他采用了由哈佛大学一位年轻的统计学天才开发的一种数学模型来搜罗球员，后来，这种模型成了专供比恩和他的下属的专属模型。

奥克兰运动家队之后的神奇表现证实了这种模型的可靠性，从那之后，这支很小的球队凭着微不足道的预算不断改写棒球赛史上的纪录，更是创造了美国棒球联赛史上最长的连续获胜纪录，仅仅一个赛季就收获了103场胜利。这是多么骄人的成绩！在美国棒球史上，也只有老牌劲旅扬基队才能与之匹敌。

奥克兰运动家队的成功在职业棒球界掀起了一场翻天覆地的革命，从此之后，越来越多的球队开始进入数字招聘行列，俱乐部愿意在“数学模型”上投入资金，从而科学地运用预测模型评估一名球员的潜力以及其市场价值。先行者大都尝到了甜头，他们比那些相对保守的同行显然更拥有优势。

如果我们不了解大数据，仅仅是看了这个棒球队的故事，所能领悟的程度大概只能到“一支棒球队令人难忘的励志传奇”，但现在你就能够明白，它实际上证实了大数据应用的一个趋势：预测性的统计分析和大数据应用，将改变传统的招聘方式，给予五百万的应聘者一个全新的综合评判机会。

在此，我们是否也可以这样想：运用大数据，组建一支强有力的优质团队将变得更加容易。

行为信息——已经构成了一座金矿

互联网以及大数据技术的发展，已经让定期获取人类的行为信息成为可能。行为信息比其他信息更有深度，所涉及的范围也更广。目前，全世界范围内有98%以上的信息都已经采用了数字化的存储方式，较2007年，这个数据量整体上已经翻了四倍。

那么，提供这些数据的都是哪些人呢？

其实，无论你在自己家中还是在公司工作，只要使用互联网，都会产生大量这类数据。比如日常工作和生活中，发送电子邮件、浏览网页、使用社交媒体工具或者从事更多的其他活动，你都产生了行为信息数据。而在产生数据时，你就是那个在无形中为发起一个全新的社会项目助力的人。

大数据影响着我们所处的时代，各个行业都已经因为大数据而发生巨大的变化。比如华尔街就因为预测股价走势的电脑程式算法而发生改变；传统的营销方式也因为互联网浏览记录的算法受到挑战。人们倾向于相信大数据对经济领域的改变，但对于人才市场，只有少数人相信类似的数据驱动的方式可以被广泛应用。

只能说，一个时代的进步总需要时间去印证，而那些走在时代前列的人，总能先于其他人成功。很多企业的HR已经把大数据工具列为人力资源管理的最得力工具。美国康奈尔大学工业与劳动关系学院教授约翰·豪斯克内西特曾对媒体说，美国国内近年来对于“劳动者分析职位”的需求大幅增长。而他自己为了配合劳动力市场供需的最新形势，把自己教学课程里的主讲科目也修改了。

那么，这是否意味着我们可以就此放心大胆地把职业生涯交给数据分析？

人们普遍不愿意做“第一个吃螃蟹”的人。在职业生涯分析中采用“预测分析法”就是那只可能会咬人的螃蟹，对大多数职场人士来说，这还是一个非常新兴的领域，有一个更贴切的叫法是“人本分析”（people analytics）。如何把理论应用到实际，这才是真正的挑战所在，况且还有很多人对此提出了道德异议。

“预测分析法”要达到预测的目的，需要建立一个全面而庞大的个人技术统计表，这其中必须包含诸如业务业绩、工作态度、技能、KPI等一切关乎个人表现的内容，其规模超乎我们的想象。

数据越真实，需要涉及的个人表现范围就越广，换句话说，这个数据就像一个私密监视器，它会窥探到人性最隐秘的地方，比如你的成长经历、生育能力、私人生活等等。所以说，大多数相关领域的公司现在也只是处于研究“螃蟹是否能吃”的阶段，他们探索的更多是此类应用的可能性。

但可以预测的是，在今后的五到十年间，数据分析行业必然会诞生新的模型，一些大规模的新实验也将问世。这对未来的企业人力资源管理以及个人的职业道路，不得不说是大有裨益的。

“选拔人才”模式的历史变迁

“公司”的概念从诞生的那一刻起，“经理人”的角色就伴随而生了，他们的任务是为公司网罗人才，在无数千差万别的人中挑选出最适合的人才，组建团队，创造效益。

关于人才选拔的历史，有一个有趣的小故事：

在20世纪初，美国费城一家制造商要大量招工，那时候有人想出了一个奇怪的方法来决定可录用的人选。这家工厂的老板让工头站在工厂门口，向围在四周的求职者抛出苹果，能够快速地接到苹果的强壮者，就能被工厂录用。

现在看来，这种方法显然没有任何科学性可言。时代变迁，如今的招聘观念和那个时代有了截然不同的评判。一些企业的精英管理者更喜欢用一种“残酷”的方式来实现优质人才目标，“优胜劣汰”成了我们这个时代的最强音。

我们可以回顾那个时代领跑的商业巨头：美国钢铁公司、杜邦和通用电气等等，它们都在用大鱼吃小鱼的方式进行整合，它们的一个动作就可能影响到整个行业的命运。弱小的企业在竞争中被吞并，由此便催生了一个更强大的企业。发明这些“残酷”竞争模式的人通常会获得业界巨头的青睐，他们被挖掘到那些更强的企业坐上更高层的职位。

一个世纪以来，这种方式都在强烈的拥簇中运行得畅行无阻。人们靠自己的卓著表现成为被选中的人才，而经理人也根据人们的表现甄选出所需要的人。正如沃顿商学院教授彼得·卡普利在论著中所写：“在预测和甄选的科学领域，没有什么方式的影响力能比得上观察人们的实际表现。”

然而，到了二战末期，美国人才供求市场开始出现一个严重的问题——人才短缺。企业里元老级的那批高层管理者年纪渐长，他们需要退位了，谁来接替他们的职位呢？必须要招聘新的训练有素的职业经理人。但这就是问题所在，从20世纪二三十年代的经济大萧条一直到二战期间，招聘就业市场始终处于疲软状态，能干的经理人就像钻石一样，始终短缺。企业急得像热锅上的蚂蚁，这时候，在普通员工中培养有潜力的人才，就成了那个时代美国商界最紧迫的任务。

从此，企业开始研究正式的招聘与管理系统。建立系统之初的参考，主要依据两个部分：一是依据人类行为学的最新研究成果，二是源于两次大战期间开发的军用技术。第一点是毫无疑问的，任何时代，行

为学都是有参考价值的；而关于第二点，则是因为爆发世界大战时国家调动了大批的军队，伤亡严重，人口紧缺，必需要尽可能高效地用人。只要你是块金子，基本上不用发光也会被发现。

到20世纪50年代为止，企业对待那些应聘专业岗位人员的选拔方法，是我们现在仍有企业在沿用的层层考核法。也就是说，花上好几天的时间，让应聘者参加一整套测试，这其中就包括初试、复试、面试等。

每个企业都期望通过严苛的测试在茫茫人海中挖掘“明日之星”，为了这个目的，过程不胜其烦。1950年的一期《商业周刊》就写道，宝洁公司在挑选高管人才的时候，会直接从大专院校入手。

现在看来，那个时代的选拔制度确实是盲目而慢效的。招聘的过程就像工厂的流水线，一个年轻人想要成为高管人员，就需要经过一系列IQ、数学、词汇、专业态度、职业兴趣、罗夏克人格、性格、健康状况等测试，最后才能在一大批“合格品”中脱颖而出，被盖上“一等品”的标签。

这些大公司的意图很明显，就是要通过种种测试考验应聘者，希望测试能够决定出合适的人选。有的员工已经被选定，也进入了工作流程，但测试评估可能还没有结束。

1956年，《商业报道》记者威廉·怀特就曾发表了一篇题为《组织人》的经典文化批评，其中指出，美国大约有四分之一的企业都在使用相似的测试评估经理和初级管理者，而这些测试通常被用于评估这些管理者是否已经准备好了迎接更高的职位。

“是应该提拔琼斯，还是搁置不用？过去，这名员工的主管们为了拿主意不得不相互讨论这个问题，如今他们可以和心理学家一起调查，看心理测试的结果怎么说。”怀特在书中这样写道。

到了1990年，这种在20世纪中叶风靡全球的企业招聘方式几乎从市场上销声匿迹了。企业放弃了花费大量时间做测试的方法，而是开始承办临时面试，面试的内容就是HR随意地提出几个问题，应试者即兴作答。

对于这种现象，彼得·卡普利说：“我认为，要是目睹现在的企业这么随意地招聘，20世纪70年代末的人力资源从业者会感到震惊。”

这个改变是因为什么呢？

卡普利为我们列举了很多原因解释这种变化。比如员工跳槽的现象增多，而测试成本太大，使企业没必要也不愿耗费资源做彻底的测试；企业更注重短期财务盈利，因而削弱了长远人才发展的内在功能；另外，1964年出台的《人权法案》也让一些有歧视倾向的招聘企业有所顾

忌，因为他们会被要求承担法律责任。种种因素的制约和影响使企业慢慢地开始把眼光投向没那么正式的量化招聘方式。持续到今日，这种方法还被众多企业青睐。

不过，上面这些因素都是客观影响，主观上是因为那些企业当时所使用的许多评估法并不科学，后来招聘方式的发展都纷纷证明了测试招聘的缺陷。最搞笑的是，有些方法并没有切实的依据，仅仅是一些从未经过测试的心理学理论而已。还有一些测试最初是为评估精神疾病而设计的，有时接受测试的人群并不具有代表性，比如大学新生，测试的结果并没有任何有意义的显示，只能显示出测试对象的反应属于“正常”还是“非正常”。

有意思的是，威廉·怀特主持了一些专门面向企业总裁的测试，测试的结果发现，没有哪位总裁的评分能够达标，他们都不在企业标准的招聘范畴。

怀特认为这些测试是不科学的，并不能客观地评估应聘者的潜能，而且有一些测试也不太人性化，对于隐私权的涉及过深，比如被测试者需要回答个人习惯或者父母的情感状况等。为此，有很多应聘者都对这些充满“攻击性”的测试题充满了极大的反感。

如今，由于数据分析的新技术和新方法，招聘作为一门学科的地位重新归位。大数据让招聘的成本更低、速度更快、覆盖范围更广、结果也更科学。不管怎么说，科技创造已经为我们带来了无尽的可能性，一个新时代正拉开帷幕。

信息采集与分析

☆信息采集与分析的重要性

对于这个世界上的大部分公司来说——我相信至少有80%的公司，哪怕它们正高举大数据的旗帜，大数据本身仍然只是一个非常空泛的概念。它们虽然十分明白其重要意义，但言行却不一致，许多奇妙的想法也难以得到彻底的实施。

对于管理上的应用和大数据的实践，企业不仅难以参与，更是不容易对进程做到完美的控制。其中一个最常见的问题就是：每天都在产生的海量数据，应该如何采集和分析？这是一个很大的困扰，就像一位经理人对我形容的："我感觉自己守着一座金山，却无从下手！"

还有一位老板抱怨："这些数据有什么用？都说有价值，但在我这里只能让我每天头疼！"

对于任何一家公司，比如电信、金融、零售业的从业者甚至政府来说，它们都需要数据来帮助自己理性决策，都有对于信息采集与分析的强烈需求。但现状并不理想，在我们国内相当长的一段时期内，比较专业的数据分析只是局限于金融和电信业。其他行业的公司对此缺乏敏感度，甚至许多从业者采取的是抵制或者漠视的态度。

尴尬的地方就在于，公司的决策者有时候更愿意相信自己的直觉，而非数据。虽然这种意识逐渐在发生变化，但有的人从来没有想过要做出根本性的改变。思维的改变从来都是艰难的，它是一块坚硬的顽石。只有当一些新兴行业开始产生，并由此崛起了一批从大数据思维中获益的公司，他们才能意识到数据的获益是如此明显。

但到这时候，大数据时代已在全球范围内到来了，这些人此时再奋起猛追，显然为时已晚。即便深刻反省并愿意付出代价弥补落下的功课，也会在一段时间内只能充当学习者和追赶者的角色。

我们都已经看到，大数据在管理领域内有望推动一场革命性的彻底改变。管理者利用大数据，可以做到很好的测量，并且做到对于数据的精确化利用，进而了解自己的公司，然后对于业务做出更好的决策。

这具体表现在我们对于信息个体的重视上，管理者必须有勇气直接将这一认识转化为改进的决策和性能。从技术的层面去重视对于数据的采集和分析，实现"数据为王"，才能改变企业落后的命运，甚至让自己成为行业的龙头。

我们知道，诸如百度、腾讯和阿里巴巴这样的中国企业已经在这样做了。但是我们希望中国的所有企业都具有这种数据收集与分析的能力，而不仅是几个典型企业。这是我们对于"大数据中国"的梦想，也是

一场与全民有关的数据管理变革的终极目标。

这几年，从我对企业的实践进行的调查来看，大数据管理的意义，并不在于你掌握了规模多大的数据信息，也不在于你的理论准备有多么充分，而在于对这些数据进行智能处理、从中分析和挖掘出有价值的信息等工作做得是否到位。假如这些实际的工作你缺位了，之前的一切准备都将失去它们的意义。

我发现大部分的公司现在还很难判断，到底哪些数据会在未来成为我的优良资产，我们需要通过什么方式将信息提炼和分析出来，转化为现实的收入。

对于这一点，即便是许多从事大数据服务的专业公司也很难给出确定的答案，人们仍在继续琢磨和进步，在大数据的浪潮中即便最好的公司也只是敲开了第一扇房门，进入了最外层的房间。但有一点是可以肯定的，在大数据时代，谁掌握了足够的数据，谁就有可能掌握未来。我们现在的数据采集，就是在为将来积累流动资产。

☆信息采集与分析的6个关键环节

信息的采集

信息采集是第一步工作，也是必不可少的第一个环节。在国外，谷歌、亚马逊等巨头早就已经开始部署数据收集的工作。在中国，淘宝、腾讯、百度等公司也已经收集并且存储了大量的用户习惯及用户消费行为数据。甚至还有专门的行业数据采集系统，而在将来，这一系统会更加发达，也会有更多专业的数据收集公司加入进来，从事信息采集服务。

信息的清理

采集以后，必须进行清理。原因就在于，当大量庞杂无序的数据收集好之后，如何将有用的数据筛选出来，完成数据的清理工作并且顺利地传递到下一个环节，是我们必须进行的一步工作。全世界已经有很多专业的数据处理公司来完成这样的工作，国内也开始出现类似的公司，比如华傲数据等。它们不断涌现，支撑起了中国的数据处理行业。

信息的存储及管理

这是一个中间环节，也是介质环节。信息进行采集清理以后，必然进行存储和管理，这两个环节是连为一体的，也是密不可分的。一般而言，我们进行数据管理的方式决定了数据的存储格式，而数据如何存储，又最终限制了我们进行数据分析的深度和广度。假如数据保存不好，管理不当，后面的一切环节都将变得毫无意义，我们将很难做到专业的分析与解读，也就无法顺利地实现大数据的实用。

信息的分析

分析信息与数据，是为了找出相关性或者因果性，进而为下一步的处理打下良好的基础。这几年来，基于开源软件基础构架Hadoop的数据分析公司呈现出了爆炸性的增长。比如康德瑞公司，它成立于2008年，是一家能够帮助企业管理和分析基于开源Hadoop产品数据的优秀数据分析公司。因为其优质的服务，这家公司仅用了不到5年的时间，市值就超过了7亿美元，其固定客户有摩根大通等著名的跨国企业。

信息的解读

解读信息就是把专业的大数据分析结果进行还原，把它体现为具体的问题，即要把“问题”转化为“问题的原因”和“问题的应用”。在这一领域内，有很多的数据还原公司提供着优质的服务，比如Wibidata公司和国内的诸多新兴的数据还原公司，都在进行这项工作，且日益专业化。

信息的量化

在最后一步，也是大数据管理的最重要环节，就是对数据进行利用。在这一环节中，通过对数据的分析和具象化，将大数据能够推导出的结论进行量化计算，然后实现它的应用。量化的目的就是得出结论，并将结论付诸实施。

对你而言，还有一个难题是：“我需要多少信息才能做出判断？”

在大多数创业公司中，管理者觉得信息越多越好，理想的情况是搜集到这个星球上所有人的数据，但在我看来，获得关键人的个人数据，对于管理的意义更大。这会让我们的大数据思维走向科学，而不是盲目地寻求海量数据。

大数据管理应用——预测和控制

对数据进行采集、管理和利用以后要干什么？我们必须做一些实际的工作，才能体现大数据的管理价值，那就是用它来进行预测和控制，而且是实现精准控制。

☆未来管理——大数据了如指掌

任何一种理念或者技术，归根结底都是为人服务的，人必须占据主导地位。也就是说，人类的需求其实是另外一个驱动，对大数据管理也是如此。假如你从事大数据管理不是为了管理人、经营人的未来，那么你就偏离了管理的本质，也就背离了利用数据的初衷。

通过大数据管理，我们可以对未来“了如指掌”。那么假如你明白了大数据管理的本质，你也就能够对自己的未来产生这种体验了。我们通过大数据，来帮助自己了解世界，进而了解人性，了解自己。一个人能达到这种境界，管理一家企业对他还是什么难事吗？简直易如反掌了。

所以，在管理上应用大数据时必须遵守的一个原则就是，要一只眼睛盯着数据，另一只眼睛盯着人。你既要在商业运营的时候知道怎么回事，明白产品如何设计、成本怎样控制等这些物理数据，又要能够洞察你的员工怎么想、在不同的阶段有什么样的需求等思维数据。

具体怎么做？办法就是把这些数据进行虚拟化。你要把它转化成数字的过程，用数字的过程来描述物理的过程、逻辑的过程、化学的过程等等。当它们虚拟化时，我们就可以非常从容地进行数据分析，来看透本质，总结和量化它的规律。最后再根据规律办事，就可以掌握未来。

☆预测——我们必须先了解自己的对象

要做出预测，你就需要先了解自己要预测的对象。实际上，大数据的应用其实是可以分成几类的，分类不会太复杂，相反十分简单。比如人、物、其他，大体就这三类。关于人的数据是什么呢？比如我们在进行搜索推荐的时候需要预测兴趣度，你要知道用户的兴趣在哪里，应该怎样去预测，这就是人的数据。物的数据是什么呢？比如对交通的统计，对天气规律的总结。理解了这些，才能做好预测工作。

现在，对于预测来说，大数据的运用已经十分广泛，其精准预测是以充分了解用户为前提的。我发现不少人都把大数据想象得十分神秘，其实这是他们不了解自己的预测对象的表现。

无论发展到什么阶段，大数据都只是一种工具而已。作为工具，它是为人服务的，而不是驾驭人的。我们在利用大数据进行预测时，最关键还是要了解自己预测的对象，并在充分了解的前提下去收集和分析与它相关的数据。

☆安全分析和管理

大数据在管理上的控制，经常体现在信息安全和危机预防上。它给信息安全带来的最大改变是，通过自动化分析处理与深度挖掘，将之前很多时候亡羊补牢式的事中、事后处理，转向了在事前自动评估预测、应急处理，让安全防护变得主动起来，而不是被动地成为一个只能善后的补偿机制。

在具体的应用中，它代表了一种未雨绸缪的趋势。我们应该切实地利用这种趋势，让自身的管理方案和大数据分析相结合，形成从数据收集分析到安全管理策略，再到效果评估的一整套安全分析和管理的解决方案，从而真正地对自己的企业管理和安全控制起到质的帮助。

捕捉问题：重点是将要发生什么

在大数据时代，人们对于数据的关注重点并不在于数据的本身，因为无论数据的规模有多大，如果不能实现数据的价值，对我们就没有任何的积极意义。有时候小规模的数据，也能为我们带来很大的知识含量并提供一定程度的价值。

所以才有人说，这个时代真正重要的，是能够从数据中获得什么，以及数据会告诉我们多少将要发生的事情。假如数据不能为你提供对将来的判断，那么不管它有多么庞大，都可以定义为“无效数据”。

前不久，美国中央情报局解封的一份材料表明，在CIA内部存在着一个名为“预测未来”的组织。他们的任务就是尽可能准确地预测未来的技术发展趋势，并且还要分析这些可能出现的新技术能否在战争中得到应用。这个部门的职能就是“预测”，通过预测来判断将来会发生什么，且要保持相当高的准确性。

而在他们的工作中，有一项任务更加令人震憾，那就是利用大数据来分析人类的行为。该部门的一名特工说：“当计算机的运算速度和存储能力大幅提高时，那么在理论上我们就可以将人类的行为编译成数据，或者我们收集到数量足够庞大的数据，然后把它们集中存诸起来，利用数学分析方法去分析这些数据的变动规律，以此来实现对于人类的群体行为的预测。”

结果精确吗？中央情报局没有公布更多的信息，但这个部门已经运行了几十年。这充分地表明，大数据在宏观管理方面的应用早已经在某些精英机构中进行，而它的主要应用方向就是用来判断将来会发生什么，而不是去记录过去和今天已经发生了什么。

无独有偶，研究这项工作的人和机构还有很多。比如麻省理工的教授埃里克就说：“我认为未来会发生一件十分有趣的事情，那就是大数据将取代人类的想法，以及代替人类去思考这个世界。”

他的假设很有道理，正是基于大数据的基本特征，也就是人们的网络浏览记录、各种数据的传输和传感设备以及GPS定位等技术的发展和社交网络数据的猛增。事实上，这种可能性已经成立了，如果我们能够做到精确地分析，就可以得出相应的结论，那就是这个世界正在发生的事情。任何一种事物都在变得越来越智能，人们做出的决定也越发合理，而这正是大数据分析与管理的功劳。

有的公司已经开始利用这一点来管理他们的部门员工，纽约有家公司就以大数据思维融入了管理制度，成立了一种预测模式，来根据过去和现在的数据精确地预测员工在未来一段时间内的工作情况。

在其他领域呢，比如金融？

国内一家银行的分行经理对我说："大数据到了这个时候，在互联网出现之后经过云计算，特别是跟网上商城、社交平台、移动终端的融合，现在又带来了数据的存储和数据的分析，这一系列的积累，最终会促成精准预测的产生。我们对于未来可以逐渐实现准确控制，通过分析数据，来判断将要发生的问题，然后把危机掐灭在萌芽状态，也就是说，大数据时代的到来和产业的发展，将使得金融危机再次发生的可能性大大降低。"

如何解决精准地预测和控制的问题？

第一，我们要有足够的数据量进行分析；

第二，我们必须找到有效的分析方法。

就像在金融领域，人们对大数据的精准问题需求最为迫切。当然，实现起来并不容易，只有通过有效的分析方法才能做到精准。比如，哪些客户是你的服务对象，你应该给客户提供什么样的服务产品更能获得满意的结果。

政府的角色

在蓬勃发展的互联网时代，大数据已经不仅仅是一架梯子，人们可以借助它摆脱旧的管理模式，走向光明的未来。而且，大数据也意味着透明化的不可避免，代表着在全球层面上，政府和国家也都要为这种透明化作出更多的努力，进行更有力的尝试，起到重要的引导作用。毕竟，相对于企业和个人来说，政府才具有推动社会进步和技术革新的决定性力量。

中国政府已经为此做了大量的工作，而且还在进一步地推出积极的举措。比如2013年，习近平主席就会见了百度公司的CEO李彦宏。对于“大数据能实现什么样的未来”这个话题，中国政府也给予了极大的关注，并询问了企业家的想法。

对于这个问题，李彦宏说：“大数据在两个方面表现出了最重要的价值，促进信息消费和关注社会民生。”除此之外，政府可扮演的角色还有社会管理，例如“热点实时监控与预警”，这包括对公共事件的监测、对公共卫生领域的提升以及对于经济数据的统计和大数据基础设施建设的推动等，使政府这双有力的手参与大数据中国的建设。无论是经济领域还是政治领域，政府的角色都至关重要，不可忽视。

其实早在2007年4月，中国政府就推出了公开政府信息的新规范，迈出了数据透明化的第一步。这一政策规定，从2008年5月1日起，政府将有关土地使用、公共卫生调查以及其他官方活动的资料都在网上进行公开，公民和企业可进行查询、监督。

现在，全球主要的大国都认识到了大数据对于国家未来发展的重要性。大数据正在成为推动企业效率提升和管理变革的强大力量，也正在成为经济持续繁荣的催化剂。正因为如此，美国政府才把大数据提到了国家战略的高度，认为这是美国继续称霸世界的一种新的战略武器，甚至比石油还要重要。

所以，大数据想在中国真正发展起来，形成全球领先的产业，政府对于大数据的战略定位是非常关键的，起着根本性的作用。政府必须抓住大数据发展的五大关键要素：基础设施、产业链、人才、技术和立法，并且要在其中担当关键的推动者和引导者的角色。尤为值得一提的是，这五个要素是普通企业所做不到的，政府就可以补上这个空白，来扮演那个提供强力支持的角色。

☆大数据的基础设施领域

一个国家在信息和存储等方面的基础设施，决定了大数据时代的海量数据能否汇集、传达、存储和应用。比如高速网络架构、大型数据中

心、商业数据托管中心和中立运营商数据中心等，这些事业投入巨大，牵涉面广，企业个人的力量无法完成，都必须由政府来牵头建设。

☆大数据的产业链布局领域

大数据的产业链并不是单一的产业链条，而是横跨了包括数据提供者、存储商、分析和挖掘商以及应用企业的诸多行业，是关乎国计民生的所有行业都在参与的划时代的“超级工程”。对于企业，往往只有应用能力，却缺乏获得、存储和分析与挖掘大数据的能力。而在这方面，当然要依靠产业链中相应的服务商，但政府在产业链建设中发挥了关键性的作用，它可以引导进行战略布局，制定相应的政策鼓励企业投向那些需要重点发展的领域，来推动产业链的快速、健康成长。

☆大数据的数据挖掘领域

政府可以设立数据挖掘和分析平台、创新平台，增强政府在公共管理领域的数据分析能力；而且还要鼓励企业设立数据分析中心，来形成一种政企联合、公私合作的大好局面。

☆大数据的提供领域：政府数据的透明化

在大数据的产业建设中，政府数据的透明化这一点至关重要，因为毕竟政府是最大的数据拥有者。但是我们都知道，让任何一个国家的政府主动开放自己的数据，都不是一件容易的事。所以这一项工作任重而道远，是一个长期的过程。但现阶段，政府可以为企业提供开放的数据平台，比如交通数据、行政资源数据等，供企业和公民根据自己的需求进行实时查询。

☆大数据的人才领域

现在，中国企业在应用大数据的过程中最缺乏的就是专业的数据人才，比如大数据工程师。培养这一类人才，往往需要投入巨大的成本，因此也需要发挥政府的引导和推动作用。

政府可以与企业以及本地高等院校开展合作，确保我们的毕业生获得必备的专业知识和技能。比如可以集中在数据分析领域开设一些本科及硕士课程，提供侧重于具体行业应用的多学科的研究方法，加速人才的培养，力争在较短的时间内积聚一批高素质的大数据人才。

☆大数据的技术领域

在大数据的存储、分析和挖掘技术与产品领域，投资往往是非常大的，尤其是初始阶段，更是一项烧钱的大工程，需要巨大的投入。但是一般的企业无法承受这样的投资，此时政府的作用就尤为重要。在技术的投资领域，政府不能缺位，且要占据主导地位，以一种长远眼光来进行技术的研发和积累，要谋在一时，功在千秋。

☆大数据的立法领域

我们知道，大数据的发展总是伴随着与个人隐私权的冲突，这是大数据产业发展无法规避的“死穴”，不能逃避，只能面对。所以，能否通过立法明确保护个人隐私权是中国的大数据产业可否获得良性发展的关键。政府应立足于防范对国内数据以及源于境外的个人资料的滥用行为，帮助公民进一步了解个人资料的使用途径；同时，在进行个人信息处理的过程中，政府也应起到加强企业与客户的信任的作用，以立法的形式规范大数据市场。

☆大数据的公共卫生领域

在公共卫生领域，由于企业可以起到的作用比较有限，更应由政府来主导这一进程。政府应该探讨公共卫生信息体系与集成应用的新模式、卫生应急数据采集技术的运用、卫生应急管理信息数据的挖掘与展示，以一种大数据的宏观思维来规范公共卫生事业的发展。要加强数据采集的技术，并保证数据采集与数据分析过程中的数据完整性、规范性、准确性、安全性和及时性，推动公共卫生事业的发展。

比如，根据收集到的数据，如果在几个地方都忽然发现某种症状的发热病人，政府这个时候就要考虑是不是某一种传染病要暴发了。这是企业做不到的，也是单凭医院自己无法应对的，需要有专业的管理部门，由政府的专业人员进行数据的分析和实时预警。

在这个过程中，当然也需要企业的参与，比如谷歌那样的公司。同时，政府也要提供相应的财政力量，根据这些数据做一些预算，通过数据分析，来决定是否需要对某些医疗卫生服务点增加资源，进行较大幅度的财政倾斜，甚至紧急动员，来预防潜在的公共卫生危机。

☆本质——政府思维模式的变革

如果说互联网重塑了我们人类的交流方式，改变了人类的文明生态的话，那么大数据则标志着社会处理信息方式的变化。属于政府的大数据时代也随之来临了，它必将引发政府的思维模式的变革。以前的不合时宜的管理方法、体制或决策应该做出改变，以适应新形式下的社会特点。

对政府而言，大数据是一笔巨大的资产。它既是经济资产，又是政治资产，还是思维资产。政府既要在宏观层面汲取它的营养，改善管治思维，也要在微观层面获得技术的进步，实现数据到知识的转化、知识到行动的过渡、行动到效果的展现。

中国怎样才能在大数据时代获取真正的竞争优势？这取决于我们的管理者是否做好了思维变革的准备。思维不变，一切不变；思维变了，一切都变。必须牢记这个原则，并把它作为大数据产业布局的指导性思想。

政府要做好准备，来迎接大数据时代的冲击，应对全社会的商业模式和管理模式的变化。对政府来说，这需要极大的智慧和勇气，去掌握数据、利用数据、开放数据，使社会每一个阶层的人受益，并借此推动全社会的进步。

CHAPTER 10 每个人的新时代：抓住大数据机遇

你的手机号码比你早一步进入数据化时代

在今天，技术的发展让我们足不出户就可以实时交流，通过手机、微信、脸书或视频电话的方式表达想法，了解这个世界。这是通信科技的飞速进步带来的文明福利，让世界进入了信息时代，缩短了时空距离，改变了我们的生活，也极大地促进了经济的飞速发展。

在数据化时代，每一种事物都在产生数据，同时也不再有数据死角。不过，并非所有人都及时意识到了这一点。许多受访者都对我说，他们不知道大数据是什么，虽然他们的手机早已经“出卖”了自己的主人，提前一步进入了这个新的世界，他们仍然对此一无所知。

☆手机智能时代——我们每天都在参与大数据

如果你清楚地知道全世界的智能手机用户都喜欢上哪些网站及有何种购物习惯、全世界的电脑用户是使用百度还是谷歌搜索及具体搜索什么的时候，你也就明白了自己的习惯。研究这些是一项有趣的工作，长期以来，我观察到人们是被自己的手机号码绑架进了数据化生活，而不是自愿进入的。移动互联捆住了每一个人，就像在人们的大脑之间系起了一根又一根传达信息的电线，只需通过一个开通网络的手机号，我们就能找到地球上的每一个人。

重要的是，其实你可以掌握这些号码的主人都在干什么。当你发现人们在互联平台不同的使用习惯之后，你意识到这是他们自己参与的结果吗？这些资源正是由一个又一个普通人贡献的，它们的价值也是无限的。

发现这一点的时候，我们就有了在数据化的世界寻求一席之地或把握更大机遇的动力。

仅在中国市场，手机的普及程度就是令人恐怖的——13亿人几乎已经达到了人手一部的程度。可以想象，每个号码都是一个接入器，也是一个源源不断地产生数据和提供信息的“车间”，只要人的大脑在思考并使用它，它就会收集和发送信息。

所以，拥有一部智能手机只是一个开始，使用智能手机的数据收集才是关键。手机永远比我们更进一步，它不但是大数据时代的小车间，同时还不断进化。每周都有新的手机上市，它的功能变得更强，平台也更方便，存储信息的空间也更大。

没错，你可能已经若有所思了：“为什么我在购买第一部手机时没有

想到呢？为什么我不能敏锐地发现它带给世界的变化？”

只是成为文明成果的享受者（包括研究如何享受得更舒适）并不能带给我们真正的成就，从中寻找到可以打开主控室大门的钥匙才是聪明之举。你应该这样想：如果掌握这些大数据的人是我，我会如何利用这些资源呢？

这时，真正的问题就来了，当数据化生活扑面而来、大到难以想象时，你是否具备分析处理它们的能力。没有人可以一口气吃成一个胖子，不过，我们至少需要提前准备一个能装下这些“美食”的盘子。

☆商机——手机使用习惯“出卖”了你

当然，在商人的狂喜背后，对于我们的消费者来说，大数据就不那么“友善”了，它更像隐藏在暗处的悄无声息的观察者和记录者。尤其对于手机用户来说，它成为了商家的免费武器。

人们用手机的行为习惯只是大数据资源中的一种，也是和我们的关系最密切的一类资源。只要你在使用手机，不管从体量的大小、信息量的增速快慢、数据间的交叉性等何种角度来看，我们都在时时刻刻地为大数据时代的发展壮大做着贡献。

手机产生的数据除了保存在自己这里，第一时间还会被各种软件，包括手机助手、安全软件、搜索引擎以及各种“云”上传到我们看不到的空间，提供给那些需要的商家（或个人），然后被用于各种分析。

从某种程度上来说，一部手机就是一个潜伏在我们生活中的信息间谍，它窥探你，然后“出卖”你。这与你使用的是苹果、黑莓或什么系统无关，只要接入互联网，你就开始了信息上传之旅。

重点不在于手机本身，而在于手机产生数据的价值。未来掌握在积极适应大数据时代的人手中。从这个层面来说，我们都迎来了一个新时代，站在同一起跑线上。区别在于，有的人第一时间就开始冲刺了，有的人还在原地左右徘徊。

透明社会，透明的机遇

一个完全的“透明社会”——就像美剧《疑犯追踪》所展示的未来那样，意味着我们的个人隐私被无限放大，并且人人都将获得某种带有前提条件的“信息自由”。后者即代表着更多的机遇和更快的信息更新速度。

扎克伯格在设计脸书的产品时，已经为自己确立了一个较为积极的理念：“现代社会正在往极端透明度的方向发展。”这理念在他的内心确立很久了，而且基于他对这个世界的观察和预测。所以在产品设计中，脸书的动态新闻、开放API、联谊会等，每一次重大的产品升级都遵循着他的这个思路。

随之而来的就是使用脸书和推特的用户很难隐藏与自己有关的信息，大家一起变得透明化了。你可以发现我的全部，我也可以发现你的一切，然后对此做些过去做不到的、不同寻常的事情。

但这重要吗？或者说，这样会带来危险的后果吗？扎克伯格说：“我认为对用户来说，这代表着机遇而非危险。”他反对那些认为信息透明会造成社会混乱与增加不安全因素的看法，认为那是“目光短浅的愚蠢之见”，应该“把他们锁在黑暗的笼子里，尝一尝孤独和信息闭塞的滋味”。

☆我们都是透明人

虽然隐私是可以设定的，为它装上一扇门，并把钥匙藏在自己的口袋里，这代表着你拒绝上网或使用电子设备；你也可以精巧地设计自己的上网习惯，始终持有保护个人信息的警惕心。但是不管你怎么努力为之，总会有一些隐私不可避免地被泄露出来，比如你对闺密的窃窃私语，说不定哪天就成为众人的饭后谈资；你对某个重要客户的评价，半小时后就传到了竞争对手的耳朵里。

如何应对这种情况？我的做法通常是，尽可能不发布会引起尴尬的隐私内容，并且公开所有个人主页，让自己率先成为一个“透明人”。我主动地让世界来了解我，反而使自己养成了发布有利信息的习惯，从而降低了有害信息的曝光机会，这对为自己塑造一个积极、透明的健康形象大有裨益。

长期透明的社会能够为我们带来三种有利的结果：

1.虚伪没有容身之地，每个人都向“表里如一”加速进化。

因为在互联网时代，你发布的信息有可能被任何人看到，你在现实生活中的活动也很有可能被人拍照在脸书或微博上发布，所以经历一段适应期（或者受到某种教训）后，人们将逐步学会为自己负责——不论

在网上还是现实中都会努力变得真实，而不是表里不一、口是心非。

这也正是大数据的目标，我们至少应该表现得真诚一些，因为数据绝不会说谎，它随时可能揭穿你的面具。比如唐骏的“学历门”事件，假如不是网络的发达，唐骏也不至于落得如此狼狈。也正是这种无比透明的社会，让以前可以轻易隐瞒的负面行为变得无处遁形。

2.社会的宽容度有所增加，虽然是一个缓慢的并不同步的过程。

由于人们保护隐私的工作日益艰巨，意味着我们某些灰色的过去曝光的机会将不可避免地增多。这时人们慢慢地发现，原来以前觉得像圣人一样的家伙也会犯错。当这种现象逐渐普遍时，社会的宽容度就会增加了。这就是一个好的结果：因为发现了彼此更多的错误，人们变得互相理解、包容。毕竟，不可能只要曝光了一件荒唐之事，就要永久剥夺一个人的社会信誉和摧毁他的正面形象。

3.形成并加强馈赠型经济，增加社会不同阶层之间的凝聚力。

馈赠型的经济是社会福利体系的升级版本，也是信息高度透明的直接产物。信息透明和快速传播以后，人们很容易关注到一些特殊现象，比如贫困群体。这样一来，发达地区的民众便可拿出自己的成果分享给不发达地区的民众——慈善或点对点捐助。在小型社区内也很明显，人们在社区论坛上发布相关消息，各取所需，成为社会经济活动的一部分。

这正是透明社会所带来的，它缩短了人与人之间的时空。相比于传统社会，每个人的影响力都在增加，人们通过分享所获得回报的可能性也在增大。可以这样说，馈赠型经济的效果已经通过社会的透明出现了，衍生出了更多的商机。这并非武断，而是一个已形成很久的事实。

例如，有人由于分享了自己对于某些事物的观点而成为了业界的意见领袖，有人则通过分享自己的内心而成为网络的新闻人物。这是由人的好奇心决定的。人们既害怕隐私泄露，又想窥视到他人的隐私，人类这种矛盾的心理恰恰是聪明人可以利用的地方，它代表着无穷的机会。

当然，是做“好事”还是“坏事”均取决于你。当你积极地对待机遇时，你从中收获到的将是理解、认同、信任和赞扬，同时也是你对于数据化生活的全新认识。直到有一天，你发现自己已成为其中的重要一员。

这就是我对于未来社会的看法。你必须先了解什么是透明社会，才能放下心中的恐惧，看到那些金子般珍贵的机遇的存在。

所以脸书的创始人扎克伯格才自豪地说：“我的目标不仅是创造一家公司，我要改变的是人们的生活，是一种社会形态。”一个透明的社会，一个信息共享的平台，通过各种工具——微博、QQ、百度、电信

还有论坛，雕琢着我们的行为方式，也改变着这个社会，让每一名普通人都不由自主地参与进来，成为它不可或缺的建设者。

☆透明社会：一切都变得简单和坦诚

社会透明的缺点在哪里？除了对于隐私的担忧以外，我们好像很难再找到其他的忧虑。

在一次讨论中，我问众人："如果说，未来的社会是透明的，任何事都很透明，你们会感到担心吗？"

在他们的反馈中，我好像没有看到他们能够找出有力的反对理由，因为透明本身是一个正面的词汇，人人都在潜意识中希望生活在一个透明社会中，来保证自己对真实信息的及时获得权。

身处这样一个环境之中，社会在整体重构和整合，人的思维也在进化。透明化的目的就是在追求一种坦诚的文化，比如信息的自由流动，大家的利益都摆在桌面上讨论，以及机会平等和避免暗箱操作（信息透明让暗箱操作越来越难）。如此一来，人们能够利用一切有效信息，针对性将更强，成功的概率也将最大化。

☆信息自由流动：忠诚和效率

没有自由就没有忠诚，这是多么简单深刻的道理。自由首先是信息的自由，只有信息可以自由流动，保证它的真实性和及时性，才能提升我们每一个人的效率。

泰普史考特在他出版于2003年的《裸露的公司》一书中早就表明了自己对于企业透明的认识，他肯定了透明化在企业中的积极作用。他认为，开放和坦诚会帮助我们降低交易的成本，减少钩心斗角，清洁办公室文化，减少管理中的政治成分，从而达到提升员工的忠诚度和提升公司效率的双重作用。

当然，假如你是一名企业管理者，我相信你已经认识到了透明化的积极价值，而且也正在采取相应的行动，来对它进行正确的管理。

观念影响速度：先改变你的头脑

德国总理默克尔说："我们要改变自己的头脑，而不是去改变别人。"

观念会影响技术进步，但更关键的是会影响我们的反应速度。当别人已在前方奔跑时，你仍然要待在后面当一个亦步亦趋的跟随者吗？就像马云说的："有些事，不做就会死。"

对任何变革来说，其本质都是一场头脑风暴，大数据也不例外，最初由技术变革开始，但最终会引起人的变革。

大数据革命，就是在革观念的命。革我们的旧观念，革我们的旧思想，同时也革我们旧习惯的命。

☆思维转变：去探求"是什么"而不是"为什么"

开拓自己思维的进化至关重要，大数据时代几乎在推翻我们过去对世界的全部认识，也从根本上改变了我们认识世界的方法——从抽样到全数据、从精确性到混杂性，从相关性与因果性的关系，都在导致人类文明的一场新变革——及观念的提升和思维的跃进。

从现在起，假如你要认识并抓住自己在大数据时代的机会，就要学着不必非要知道现象背后的原因而去分析现象，因为你只须了解现象之间的相互关系并利用这种相关性来挖掘因果性，即可达到最终的目的。就好比你在两个点之间直接划出了一条直线，跨越了一切曲线障碍。

同时，因果关系仍然是重要的，但已经不再是现象的直接原因，而是一种"结果"。我们是在探寻相关性的时候发现了因果性，而不是像以往那样通过探寻因果关系来了解这个世界了。

这里的问题是，因果关系并非变得无足轻重，而是不那么浅显易见，所以我们放弃过去事倍功半的做法，转而通过简单和易于理解的数据来挖掘本质。单纯探究因果往往会导致你对世界的误解，所以为了警惕再次犯下此类错误，你要建立的新观念，就是学会让数据发声。

☆量化思维：抓住高价值机遇

对于今天的企业而言（不论大企业还是中小企业），都面临一个全新的重要挑战：如何才能甄别和分析影响客户满意度、导致高价值客户流失的关键指标。这是一个普遍问题，模糊思维只能产生对机遇的粗略判断，很容易出现关键错误，量化思维才能让一切变得清晰明朗，在精确和全面的数据基础上，对关键目标设定量化的标准，修正过去的错误。

大数据的量化思维：

● 制定详细的蓝图：目标明确。

- 在一开始就以客户为中心：结果明确。
- 设定优先业务及确立盈利策略：业务明确。
- 从现有机遇入手获取阶段性成果：计划明确。
- 根据反馈的结果创建纠错机制：监督明确。

总而言之，随着技术的进步，我们需要提升洞察力，在竞争中首先创建智力优势，才能最终体现技术的成果。大数据思维在本质上就是一种分析和洞察思维，所有的核心工作都围绕这一点进行，它提升了我们的认知，优化了实践的方法。学会利用大数据的思维去分析和洞察世界，我们就能得以持续地清醒。

新的成功模式——赢在利用数据的能力

大数据时代在中国的到来，对很多企业来说都是一个契机，但同时也意味着全新的挑战。对个人来讲，如何应用一种模式成功地管理这些数据，就需要我们重新审视。有人提出一个观点：忘记“大数据”，尝试从“全数据”的角度思考。

这就是一个换位思考的问题，一旦企业管理者弄清楚了角度问题，就可以更加专注地分析和研究如何运用这些数据。

因为在数据采集和数据挖掘之间还有一个要解决的问题，那就是认清两者的差距。它经常导致我们损失几百万甚至上亿元的利益——现金或商业机会。

大数据根本目标是体现数据的既有价值。

中国的大数据先行者已走在了最前面，比如淘宝、京东等电商企业，无论是软件、硬件还是服务，它们都已经积累了无可匹敌的实战经验。它们针对普通消费群体的解决方案已在市场上取得了优秀的成绩，做到了把数据的价值最大化。

从经济学角度看，大数据更像是一种全新的营销方式。正是这一点决定了我们每个人都有机会成为这一大转型中的弄潮儿。大数据本身并不意味着某种全新的技术被发明出来，它不是技术创新，而是技术整合——为了让数据的价值得到体现，把已经有的技术按照应用需求进行全面提升，如同我们打通了所有房间并让它们连为一体。

如果用一个形象的比喻，就像你开了一家超市。起初，你只是把货架放在那里，所有的物品都摆在上面，任由顾客进来自行挑选，这样效率很低，因为顾客并不清楚他要买的东西放在哪里，必须自己挨个儿去找。怎么整合呢？你要对已有的商品信息进行分析后作出最合理、最利于营销的方案。比如啤酒应摆在哪个位置，才能让顾客一眼看见，实现销量的最大化；香烟摆在哪个地方，才方便顾客购买。

假如你正确分析了顾客的习惯，你就会发现香烟放在结账的地方最利于销售，因为吸烟的人多为男性，他们不习惯到超市里面四处张望，而只为了买盒烟。人们的潜意识中认为吸烟是非常不好的，所以他们盼望买烟这一程序简单化。

另外，如果你分析了顾客的思维信息，你可能惊讶地发现：啤酒摆放在婴儿纸尿裤的旁边最为理想。我们在前面已经提到过这一事实，那这是为什么呢？因为被“命令”前来购买婴儿纸尿裤的一般都是婴儿的爸爸，他们来干这活是“迫于无奈”，心情十分不爽，所以看到旁边的啤酒，许多人会顺手买走一两瓶。这就是通过整合数据、分析信息来获得

价值的体现。

如果你自己创业或作为企业的管理者，会如何收集和利用手中的数据呢？它们可是让人流口水的财富。像一家房产中介公司，你要通过不断地分析社会经济景气指数，来了解社会的整体经济运行情况，判断未来的房地产市场会发生怎样的变化，民众的消费和投资理念会如何转变，从而调整你的战略，让手中掌握的数据价值最大化。

事实上，几乎所有的人（参与数据产业者），不论是中介公司、政府机构或市场调查公司，长期以来都在对自己掌握的数据进行整合分析。这可能是一个被动的无意识的工作，以前叫作数据挖掘而不是大数据。这说明数据的价值老早就被人们认识到了，并已展开了相关的投入工作。人们要在众多收集而来的数据中去分析用户的喜好，然后为自己的业务决策进行指导，并得出相关的正确结论。

耐克公司的新产品有一个绝妙的好故事。它是运动产品领域内第一家嗅到并抓住大数据商机的超级公司，比如Nike+跑步鞋。早在2006年，耐克公司就一直在寻求如何让跑步变得有趣，最后发现可以结合音乐来实现这一目标，让本来枯燥、乏味和耗费体力的跑步活动重新充满乐趣。于是，Nike+iPod诞生了。

在第一阶段，Nike+只是完成了“一边跑步，一边听音乐”。这种事，事实上许多公司也在做，但耐克公司的突破在于，他们的研发人员开始盯上数据的作用。比如，通过在鞋内加上传感器，再给iPod装上接收器，这样跑步者就可以实时地看到自己的步速、距离等一系列的跑步数据。

第二阶段是一个更高的目标，即利用大数据构建起来的互联网社区，人们可以分享自己的运动数据，与社交结合起来，成功地将运动与生活完美融合。如此一来，人们不会把这样的鞋穿两个月就扔掉，而是与耐克公司建立了更加牢固的关系。耐克不但让消费者长期购买、使用自己的产品，还得到了有益的信息反馈。

相关财报显示，2013年第一财季，耐克公司在北美的营收达到27亿美元，同比增长高达23%，由于对大数据的及时介入，公司的业绩飞速增长。

到现在为止，技术水平的突飞猛进使得数据的价值更加突出，从而改变了成功的模式。谁更擅长分析和利用数据，谁就更容易成功。一个行业必须不断创造新的需求才能赢得更多的市场，分析和利用数据则能够实现这个目标。我们利用信息技术进行数据挖掘，可以让海量的信息创造出更多更实用的价值。

数据的收集可以通过互联网完成，跟踪和统计都十分便捷，有时甚

至不需要任何人工，有一台电脑就能够做到了。例如，软件可轻松地实现对数据用户使用痕迹的跟踪。比如，消费者所有的上网行为、浏览的网页，相关运营商把这些信息自动收集，提供给需要的企业或个人，然后通过这些数据来判断用户的喜好，进行针对性的营销，相关各方都从中获取了价值。

总的来说，如果你要在某一方面取得成功，收集到的数据当然越多越好。但更重要的是，你需要学会运用它们，让这些数据去创造和提升客户的体验，帮助你开拓和巩固市场。在所有的关键环节，你都要将数据充分地应用起来，才能助你缩小与先行者的差距，并与市场展开互动。这是我们了解市场的窗口。

删除——哪些数据是危险的？

从大数据时代开启的第一天起，我们的生活就注定被数据灌满了。这是一个人人都需要隐私但又不懂得在乎和保护隐私的时代，几乎所有人都在发布数据，把它们挂在网上或传播到公共平台。人们既向外发散，又向内吸收，自觉或不自觉地收集各种各样的数据信息。这时，如果你不学会聪明地“删除”它们，你的生活就成了无用信息的垃圾场。这就是删除数据的必要性，同时我还想告诉你一句话：既要利用数据，又要忘掉数据。

对于我们而言，遗忘才是常态，记忆只是一个例外。为什么这么说呢？因为在每天产生的所有数据中，几乎大多数对自己都是没有价值的（虽然对别人可能意味着不可估量的价值）。你接到30条短信，其中27条是广告，要删除它们；你读了一张报纸，只有两条新闻是你有兴趣的，其他的也要删除；你看了一场电影，看完后觉得索然无味，后悔不已，那么也要尽快删除这次不愉快的体验。数字技术的发展和网络力量的壮大，让数据充斥了我们的生活。充分地利用它们可以帮助你成功，但能否清楚地辨别它们，则决定你能取得多大的成功。

在数据像病毒一样扩散的时候，记住那些对你有用的知识，删掉那些对你无用甚至有害的信息。有时候这可能需要很高的成本，就像每家公司都要购买一台碎纸机，来毁灭那些“无用”和“有害”的文件一样，你也需要一台数据碎纸机，把它列进你的日程、装进你的大脑。否则，如果你一字不漏地记住了一切，这不仅令人发狂，而且使人绝望。

这道工序就像在邀请你参加一场辩论会。在这场持久的辩论中，你去寻找逐条铺开的详细论据，准备标出那些你认同的地方，把矛盾之处标出并认真留存，思考可能获得答案的东西。然后呢？不值得付出精力的问题或部分则跳过去，把空间留给后面的议题。

这就像必须定期清理电脑一样。人们在电脑中不断地建立新的文档，存入新的信息，同时每过一段时间，也要清除那些过期和失去意义的文档。它们可能是几部已读六七遍的网络小说，几部已对情节滚瓜烂熟的电影，几十张图片，几千个上网记录或操作电脑留下的其他垃圾文件。

这些数据不但已无用，留下来反而是危险的，十分可能被别人收集整理并分析你的生活习惯，比如像那些黑客正在干的事情。因此，我一直认为，人们在理解大数据之前，就积极地拥抱这个强大的工具，可能是一个巨大的错误。假如你不懂得保护或删除个人隐私的话，只有在它们泄露出去的时候，你才会体会到这是多么令人惊慌而又恐惧的事情。

对于大数据狂热的网民们，我的忠告就是牢牢抓住私人信息的绝对控制权，学习如何避免它们落入不恰当的人之手。

假如你已被数据压得喘不过气来，那么从现在开始，对它们进行适当“删除”。我们必须使自己的生活只是生活，而不是别人眼中一览无余的展示品；我们必须学会分辨然后做出果断的决择，让数据经过一道筛选的工序，留下有营养的，去掉有害的。就是这么简单，但让每个人都意识到并采取行动，却并不容易。

现在，重要的是预见未来

大数据对个人最大的作用，是我们可以通过它预见未来，发现哪些是优先信息，哪些又是次要和不重要的信息。这就像一个人走在漆黑的路上，突然看见了一个发着亮光的指示牌。这个简单的指示牌上面写着：

往左，是去长途汽车站的柏油公路；

往右，是去飞机场的地铁站；

往前，你将看到一间便捷酒店；

往后，是你来的方向，但有一家你忽视了的美味家常菜馆。

一个很小的指示牌，清晰明了地让你看到了自己的未来——今晚我将何去何从。这就是大数据的作用。通过精密系统的预测，我们的未来很容易被别人掌握，但我们也可以通过这种工具来发现这一点，帮助自己做出有效的决策，而不是在黑暗中抽着闷烟不知道该如何去做。

未来总是比过去具有更高的价值，谁掌握了未来，谁就掌握了这个世界。有人说："在未来，我们每一个人都有15分钟的成名机会。"这是一只无形的手在左右这15分钟。这只手就叫作"信息"。比如，在信息饥渴的时代，谁能提前15分钟获知一条股市内幕消息，谁就是下一个一夜暴富的人，那些哪怕只晚了一分钟的人就不可能得到这么多的财富。

成功的机遇总是转瞬即逝，谁先看到，谁就是赢家！在这个世界，人类很多行为都遵循着一些既定的也是固定的规律，一些事情的发生总是有章可循，人类90%的行为是可以预测的。那么，如果我们可以提前一些时间根据一些信息预见到这一点呢？结果就是你成功了。

也就是说，当你将生活数字化、公式化以及模型化的时候，把它们放到大数据系统中进行分析预测，你就会发现其实每一个人都非常相似，每一件事的发生和演变也都十分相像，而且十分有规律。虽然有些事情看上去随意而且偶然，但极容易被预测，只要你发现了不同信息之间的巧妙关联。

现在，各行各业的机会都特别多，到处都是创业的热潮，充满诱人的机遇。但事实上，却有无数的让人看不清楚的地方。大多数时候，人们只看到了一个成功的例子，却根本没有发现那些倒在后面的失败者。

你身边有成千上万的人在向前冲，试图超越你，而你也希望超越别人。你应该怎么做呢？如何预见到未来，规避风险呢？这就是大数据能给你的东西，也是你可以把握的东西，前提是你必须认真对待它，就像成熟地对自己的生活一样。

大数据既是人类社会的行为统计学，同时还是一门可以概括宇宙运

行的科学，它致力于统计和归类每一条信息，只要你能针对它设计一套适用于自己的系统，灵活实用地收集和利用数据，你就开启了属于自己的新时代，完全告别过去，并且全新地走向未来。

CHAPTER 11 掌握大数据，做未来世界的主人！

正视现实——无所不在的眼睛

如果你某一天醒来，突然发现自己的世界是透明的，不要感到惊讶，因为这是大数据发展的必然。“棱镜门”事件的曝光已经开始让人们意识到：数据没有做不到，只有你想不到。随着大数据技术的渗透，不但相关产业得以蓬勃发展，而且某些机构甚至个人也拥有了无所不在的眼睛。他们通过这双眼睛，既能够透过表面信息看见深度信息，也可以透过过去和现代的信息发现未来，然后控制这个世界。

这就是现实。要掌握大数据，你就必须先面对这个现实。大数据既是企业的“杀手锏”，是个人的“月光宝盒”，同时也是一把“双刃剑”。从定义上看，大数据是“在各种各样的数据中，快速获取信息的能力”，它最强调的其实并不是“大”，而是数据的多样性、处理的速度和获取价值的广度。这也意味着，我们获取这样的价值，就会付出相应的风险。

美国的奥巴马政府早在两年前就将大数据战略上升为美国的最高国策，认为大数据是“未来的新石油”。那么与之相伴的，就是美国政府将对数据的占有和控制作为陆权、海权、空权之外的另一种国家核心能力。

结果就是：“棱镜门”计划成为美国的大数据战略的一部分。所以才会有军事专家戏言：“损失了一个斯诺登，相当于全球最强大的美军损失了足有10个装甲师的兵力。”因为斯诺登手中掌握的数据要是被其他国家所利用，就会对美国的安全构成极其严重的威胁。数据已不仅是价值生产品，还成为了一门可摧毁一国经济、金融、信誉，甚至国民凝聚力的超级武器。

反过来，这也意味着数据对个体的极端重要性。在第三只眼的注视下，个人的隐私已无处遁形，各种各样的风险在连续发生。看起来，好像不是我们控制了数据，而是数据控制了我们。所以作为政府、企业以及个人，都要进行清楚的定位，明白自己在多大程度上可以或者采用什么样的方式来使用获得的数据，以避免透明社会产生的副作用。

☆威力无所不在——悬在我们头顶的全息镜头

普通人对大数据的认识是一种完全陌生的状态，似乎相距甚远，但它的威力已全方位展现，并已经渗透进人们的生活。即便在睡梦中，也在被进行数据收集。比如，家中电量的使用情况，被电表随时记录并在电力系统内实时追踪；信用卡系统在记录你的消费信息；能源公司在查

看你的燃气使用情况；交通系统在整理你的违章记录。还有什么是被遗漏的吗？除非你脱离现代社会，到深山老林找一处洞穴隐居，也不使用银行卡和电话，否则你就会被数据收集方定位并且归类分析。

☆警惕个人数据被无休止滥用

人们走到哪儿都会被收集“脚印”，这些个人数据被拿去分析，据此，政府或商家提供优质的个性化服务。这当然是大数据应用程序的魅力所在。虽然许多公司强调收集、储存、分析数据都是“匿名”的，不会泄露这些数据，也不会进行滥用，但事实可能并非如此。

普林斯顿大学的电脑专家迈克先生在一次论坛上表达了他的担心：“可供分析的数据越多，就越不可能保持匿名，要识别一个人其实只需要几十个字节的信息量。”这意味着，个人数据被随意泄露或向更多的非必要知情方提供权限，已是一个不公开的事实。

这当然需要警惕，我们如今面临的一个极为迫切的问题就是：“我是否真的愿意接受一个由数据系统控制的世界，哪怕它正在一天24小时不停地监控我？”数据化生活为我们打理好一切，方便人们的生活和工作，甚至有助于光棍提高相亲成功率。这是惊喜，但它的背后则是“数据暴政”。它观察并记录了我们的每一秒钟，每一个想法，每一次行为，却丝毫不在我们的可控范围之内，这是无法忽略的现实危险。

再好的事物都像双刃剑一样，既有好的功能，亦有坏的副作用。“棱镜门”事件提供的教训已足够让人畏惧，因此，在数据的世界中，人人都是不安全的。

一旦这些数据被“有企图”的人使用，就会立刻变成一把杀人不见血的匕首，随时可能毁灭我们的幸福生活。

规避风险——让数据控制一切

在这样一场全民参与的“数据革命”中，各行各业都在发生深刻的改变，包括我们的思维方式。但与此同时，也引发了人们对于“数据暴政”的担忧。这同样是一个值得警惕的问题，尤其对于我们个人来说，怎样避免它的风险，并成为这个新世界的主人？

数据不管有多少，海量或是无限量，它仍然有一个极限点，无法到达也不可能统计出来。假如你将命运完全交给数据，它在量化你的同时，也会成为你命运的暴君。完全根据数据办事带来的负面效应，正是今天许多大数据学者所忧虑的，因为数据缺乏最珍贵的人性。人的判断和人性的特点，是枯燥和单一的数据不可能表现出来的。

比如一部上映于2002年的美国科幻大片《关键报告》，讲述了在未来技术先进到了警察可以阻止犯罪发生的程度——通过海量信息分析，在嫌疑人还没有犯罪之前就把他拘捕，以预防犯罪。至于怎么知道谁要犯罪，则由三个躺在水池里具有特异功能的人决定，他们能及时捕捉一些关键信息。这有点像美剧《疑犯追踪》中所描述的，数据告诉我们一个人即将做些什么事情。

但事实是，毕竟还没有发生，不是吗？人性最大的特点就在于不可预测性——预谋好的犯罪有可能终止；准备好的计划有可能终结；想做的事情突然不想做了……诸如此类，我们每个人都有很多这种临时放弃某件事的经历，这是数据意识不到的，也是无法判明的。

另一个悲情的例子是，华尔街证券市场曾经通过复杂的数学逻辑设计出来一套交易策略，但最终酿成了市场崩盘的结果。

美国科学作家莱特说：“科学的数据与对人的数据总有很大的区别，像天文、气象、传染病的研究资料，是经过科学家精心收集实验所得，它们是宝贵的资料；但对于人的研究资料，正像我们对人性所了解的那样，是多变和可逆转的，所以总是不太可靠。”

☆数据的“风险管理”——放到一个地方是危险的

在对数据进行管理的时候，要恪守一条定律：把全部数据都放在一个地方将承担最大的风险。就像你将自己所有的钱都存进了一张银行卡，而你又经常用这张卡在网上进行购物，也没有为自己的网银提供足够保护的话，你的这些钱随时可能不翼而飞。钱不能放在一个篮子，数据当然也是如此。

为了资金的安全，我们应该分开存储，而不是押宝一个篮子足够安全。比如，有些电话号码不能存进手机而是应放在秘密的本子里；身份证和银行卡要分别存放而不是放在同一个抽屉中。这有利于分散风险，

假如你真的遭遇到了窃贼，你就明白这是多么重要。

对企业而言数据的安全性更为重要，像数值数据可以存储在数据库里，非结构化的数据则可以存储在文档或者表格里，进行针对性的管理，来分散风险的来源。我经常看到一些企业犯下低级的错误，他们在做完架构、设计、开发等所有的工作之后，才开始考虑安全问题，就像吃完了饭才发觉这顿饭是凉的，可能会伤肚子。这是非常大的错误，不能杜绝危险的发生，只能起到事后追补的效果。企业应该在开始之初就考虑数据的安全问题，来搭建安全的架构，对数据进行严密保护。

为数据建设一个铜墙铁壁般的房子只是安全的一个方面，为了保证数据的安全，企业还应该将数据切片进行存储，以此做到更为精确的控制。什么叫切片存储呢？就是只对单名员工开放部分数据的权限，只有两人或多人以上，才能查看到某一部分完整的数据。如此一来，就算有人侵入数据库盗用了这个部分，总体还是安全的，因为单一部分很难获得全部的信息，甚至有时一点作用都没有。没有上下文的数据对于窃取者来说可能意义不大，特别是当数据的价值密度很低时。

☆加密——消灭数据的“毒性”

有毒的数据我们称之为“毒数据”，这个词由费里斯特提出，被称为toxic data，意指企业手中掌握到的如果泄露出去就会对企业或个人造成巨大损失的数据。比如电信公司收集到的数据，其中会包括用户的通话时间、地点、移动轨迹等；社交网站收集到的数据包括用户的登录密码、发言和好友信息等；金融系统收集到的数据，则包括用户的消费记录或消费习惯等，用户的银行密码当然也包括在内了。

为了降低泄露这些数据的风险，对于它的加密就变得尤为关键。也就是说，数据必须被锁在一个完全可控的保险箱内，确保我们每个人都能成为这些信息的主人，而不是在它被泄露时无能为力，任由它成为一种“有毒物质”。

现在，大数据领域内最基本的做法是使用透明数据加密法，这一做法代表着对所有的捕获到的数据都进行加密，以此保证全部数据都具有同样高的安全性。虽然它的成本一度很高，但近几年来已逐渐变得可被中小企业接受。

另一方面，如果我们的生活全部由数据控制，甚至包括你的思想，会引发多么可怕的后果呢？现在，数据正以亦好亦坏的方式控制我们，而且已不断地证明人们比想象中还容易受到它的驾驭。

在它的控制之下，我们的生活正发生怎样的变化？

1.不受制约的数据收集，正大大地威胁到我们的隐私和自由，这是显而易见的一个负面作用。

2.数据控制一切还加剧了一个早已存在的风险：人们正越来越依赖数据，但它远远没有我们意想中的可靠。

当依靠数据的分析并不完全可靠时，我们可能会完全受限于分析的结果。一个错误的结果，却不会受到任何质疑，甚至还是堂而皇之地持久成为某种权威结论，继续加深人们对于数据的依赖和痴迷。最终，人们可能仅仅为了收集数据而去收集数据，或者赋予它根本无权得到的信任。

有一位经理人说："我现在离开数据，就无法做出决策了，我像相信上帝一样相信它，除了上帝，任何人在我面前都必须用数据说话。"这是他的信仰，同时也是很多管理者和决策者所遵守的原则，所以这句话经常回荡在华尔街、中关村或者上海浦东的高档写字楼里。在我看来，他们都在为数据打工了，已成为数据的奴仆，让它控制了自己的灵魂与思考。

长此以往，后果将不堪设想。

最后，我们如何避免数据的这种独裁和垄断，也就是数据主宰一切的困局？我们怎样与它平等对话，灵活协商，而不是没有条件地言听计从？

在我看来，摆脱数据独裁的唯一办法，就是建立起一个可以持续的数据协商制度，这意味着把分析和使用数据的权力给予基层员工，让那些有一线经验的员工来判断信息的来源是否正确，来预见数据的分析是否合乎情理。

就像大数据的产生是由于技术的分工整合一样，使用数据的人也应该更好地分工协作，集思广益，比如让一群经过先进技术训练的数据专家共同来解决棘手的难题，让这些数据使用者承担起更多的责任，避免出现数据的独裁困局。

结论是：我们必须让数据说话，但是"钥匙"一定要掌握在人的手中。

越过障碍——流动性与可获取性

在实际的应用过程中，数据的“流动性”和“可获取性”是一个必然的障碍，就像财富向有钱人集中一样，当数据变得越来越有价值时（海量资产），就成为了一个被追逐和被垄断的。

人们一方面渴望获得更多的数据，另一方面，则面临高价值数据被垄断的障碍。

由于大数据带来的挑战是跨行业和跨领域的，所以数据在不同的行业和领域之间的流动性非常重要。数据若不能顺利流动，大数据便不能开展；数据无法被获取，大数据分析也就失去了前提。

云计算和大数据的兴起，注定会在数据公开领域带来一场革命，无论是对社会、公司还是个人来说，都是一次对信息的世界观的改变。这意味着数据不再是自己的不可展示的私有产品，而是融入了生产方式，成为可用来交换的资产或者增值工具，也变成了竞争和生存的关键。

就像在工业革命时代，人们都需要用到电；在计算机时代，人们又都需要电脑；那么到了大数据时代，我们人人都需要自由地与合法地获取数据。它带来的是竞争形态的改变，同时也是竞争思维的变革——我们既要保证自己及时获取数据，也要尊重并支持别人的相关权利，互相满足对数据的需求。

比如在2009年，美国政府就创建了Data.gov网站，为大数据的普及和数据的公开敞开了大门，公众能够通过这个网站获得各种各样的政府数据。中国要赶上大数据的变革，首先要开始一场深层次的“数据公开”行动，从政府开始公开数据，其次是企业，最后到我们每一个人。

数据控制原则一：确保数据有最大的可获取性。

数据控制原则二：确保数据有最大的可理解性。

我们都知道数据中蕴藏着金矿，但从基因组学、天文学、生态学、临床医学到高能物理等，正如上面我讲到的，这里的核心问题是，当数据像洪水一般涌来，我们如何进行数据的收集、管理，确保它的可理解性和可获得性？

大数据的复杂就在于它交付和使用的速度，比如一定要实时，如果滞后一两个小时甚至一两天，它可能就失去了获取的意义。实时流动才具有最大的价值。所以，数据的流动性是大数据实现其个性化应用的最大基础，换言之，数据本身没有价值，有了足够好的流动性，它才具有了价值。

例如，在美军驻阿富汗的某座基地的电脑中，存储着与一伙恐怖分子有关的信息，包括他们照片、武器装备和后勤等几乎所有的细节，这

时一架无人机在山区发现了一群不明身份的人，需要确认是否就是打击目标。那么，基地就需要通过数据链实时将这些人的原始数据发送到无人机，供它在空中进行对比，如果不能实时获取，一个小时后，这伙恐怖分子可能就跑远了，或者无人机已经没有燃料了，只好返航。最坏的结果也可能是无人机向下面的人发射了导弹，几天后才发现误炸了平民。这就是流动性与可获取性的典型案例。

事实就是，我们周围的这些数据，在自己不能够用起来，即数据没有流动性的时候，是不具备什么实用价值的，只有排除掉流动障碍，可以进行个性化的按需获取，它才真正具备了大数据时代的特征。

当然，我深信数据的自由与合法流动一定会到来，这将是一个无法逆转的趋势，也不会由人为干涉来决定。在这个推动的过程中，会有源源不断的人站出来，把自己的技术处理能力和处理方式提供给更多的人，而且是以相对合理和低廉的方式提供，然后共同推动数据的流动，完成个性化应用。

避开死角——错误的前提会导致错误的结论

为什么我会说错误的前提导致错误的结论？数据分析依据是否必须由人的某些动机来决定？那些每天都与数据为伴的人或许可以用一句话来概括它的原因，但刚开始认识与接触大数据的人可能感到困惑：既然数据的相关性已足够体现某种“事实”，为什么还要为数据的分析设置前提呢？

答案可能是很多人不想接受的，因为他们会发现自己已经得出了太多错误结论。有时候，是由于拿来分析的数据质量不佳或数量过少；不过多数情况下，恰恰是因为我们误用了数据的分析结果。我们自己的错误让数据分析出现了错误。

当人自身出现问题时，大数据要么会让这些问题继续存在，要么就会加剧这些问题导致的不良后果，使结果向错误的方向越走越远。

大数据技术生产出来的“数据”，不一定就等同于好的数据。你一定要先明白这一点并对此做出清醒的判断，否则就将陷入盲信和迷信的泥潭。现在已有越来越多的专家坚信我的分析，那就是大数据并不会自动产生好的分析结果，而是依赖于你给它提前设置的条件，比如某种分析逻辑或者数据的侧重点。

在具体的运用中，假如数据不完整、断章取义或者被破坏，可能会导致我们产生错误的决策。甚至从某种程度而言，这种灾难性的结果是一定会发生的，从而削弱数据的价值，影响企业的竞争力或者我们个人的日常生活。

格林先生是美国哈佛大学的教授，同时也是定量分析领域的专家，他就曾经因为在进行数据分析的工作时做出了错误的理解，导致结果谬之千里。他在过去几年发起了一个与大数据有关的分析项目，工作内容是通过检测推特和其他的社交媒体帖子中的“工作”“失业”和“分类”等关键词，来预测美国的失业率。

他的工作小组通过情感分析技术，收集了包含这些关键字的海量内容，根据这些帖子的增加或减少来判断它们与每月失业率之间的相关性。在收集和分析过程中，小组成员发现包含关键字“工作”的内容急剧增加，也就是说在某一个月有更多的人在讨论工作话题。但随后，他们发现这与失业率并无关系，真实的情况是乔布斯去世了——乔布斯的名字Jobs也含有“工作”的意思。

格林因此说，人们应从这个例子中吸取教训，不要完全相信大数据可以在没有任何条件的情况下告知你一件事情的结论，并神奇地帮助你做出决策。所有的分析都必须设置一个靠谱的或精确的前提，否则就可

能把你的结论引向与事实毫不相干的地方。

也就是说，在缺乏必要因果关系支持时，数据之间的相关性可能会给你带来灾难性的失败。解决这一麻烦的方法有很多，比如我们可以通过添加额外的关键字来增加分析前提，但往往也需要大量的人力工作。

在设定某些固定关键词时，起初我们会从数据的分析中看到一些相关或者无关的东西，相关的多一些，无关的仿佛真的很少。但随着时间的增加，如果你不更改查询，不修正前提和数据背景，你会发现含有这些关键词的话题正以某种方式逐渐偏离主题。某些时段，它们偏离较小，但有时候却非常大，让你几乎找不到它们之间有什么关联。

不过格林也承认："总体而言，很多大数据分析都产生了有用的内容。重要的是，我们只要为分析工作设置必要的启动程序，引导它在一条正确的轨道上，它会给你计划中的结果，完成传统方法做不到的任务。"

数据本身并不等于智慧，只有经过正确分析之后，数据才能凸显它的意义。如果人们觉得大量数据能够奇迹般地产生良好的分析结果，而不需要人工任何干预，那么它消极方面的问题可能就会走上前台，阻碍我们做出积极的判断。

乔布斯的名字是一个经典的案例，在他去世时（该信息的背景和前提发生了变化），同一个关键词对于数据分析的结果就造成了极大的干扰，把终点引向了与出发点风马牛不相及的地方。

《华尔街日报》的一篇报道也认为，今天有越来越多没有内容的数据在推动人们的决策过程。但真相是，并非数据无用，而是人们利用数据的动机发生了微妙的变化，就像在炒菜时放错了调料一样，尽管只是一丁点的错误，菜的味道就完全改变了。

对没有设置正确前提的"相关性"不利一面的分析始终是大数据研究的热点，比利时大数据专家费兰克在他的一篇文章中指出，在某些情况下，银行会因为用户在社交网站上的联系人的情况而拒绝给用户贷款。虽然这个人的信用良好，但他有一些喜欢赖账的朋友，因而影响到了银行对他的判断。"相关性"在这里就伤害了一位原本有资格获得银行贷款的公民。

这表明，当我们不经任何前提而直接从数据的相关性得出结论时，必须进一步分析，否则就可能带来麻烦。比如美国20世纪的一些刑事数据显示，驾驶入门级豪华车且年龄在20和27岁之间的西班牙裔和黑人男性最有可能是毒贩。但在警察实际办案的过程中，却发现许多合乎该数据条件的非裔美国人并不是犯罪分子，而是良好公民。他们中的许多人被警方列入了重点监视对象，可最后虚惊一场。

简言之，大数据是一个分析工具，但不应该被我们视为不论何种情况都始终正确的解决方案。它可以帮助你缩小范围，从数百万可能缩小到150左右。但是，即便岁月再过200年，我们也不可能去将“判断一切”的机会交给电脑。我们不能只是依靠数据进行分析，不能忽略人类的智能在分析过程中起到的独特判断力。

如果你这么做了，一定会给你带来难以摆脱的烦恼。到时候，大数据在你的生活中就变成了一个致命的大麻烦。我的一些朋友已经体会到了这一点，而我希望人们不再犯下此类错误。

解决问题——定位人的角色

现在，通过全书“不厌其烦”或“有所选择”地展示数据在今天多个领域的应用，我们已经非常清晰地理解到了大数据时代的内涵，它是一个具备海量数据被共享或被搜集、追求相关性、不再迷信采样调查而是追求整体分析的时代。

——它的基础是人类经过几百年发展的卓越科技基础与铺天盖地的网络平台。

——它让我们几乎没有秘密可言，这是科技赋予它的权利，也是科技塑造人或人塑造科技的选择路口。

——它让电脑越来越“聪明”，甚至可以筛选更加适合自己的模式或信息，自动地帮助自己改善它的运行模块，虽然它还不能统治人类。

——它用海量数据弥补了个例精确性的不足，然后导向更加精确的结果。

——它产生了相关性和因果性的辩证关系，数据加工者根据这两种关系的不同选择，在预测人们的行为、疾病的发生和灾难的到来时也会产生不同的结果。

——它不可避免地导致了商业模式、政治格局的变革。

——它赋予了使用者庞大的权限，但是这种无孔不入的权限让人感到恐惧，甚至会引发更严重的对于人类现有文明秩序的威胁。

——它改变了法学思想，在司法领域产生了一个关于无罪推定与有罪预测的深刻问题，比如美剧《疑犯追踪》所展现的。

在大数据时代，我们每天都在暴露出巨量的个人信息，它的巨大价值在于二次利用，而这是我们目前暂无法监管与救赎的层面。如何保护必要的个人隐私并成功地阻止大数据巨头的收集，是每个人都在讨论的紧迫话题。

我们从本书的收获或许当然并不限于上述种种，而在于对中国人世界观的拓展，也是对数据和人的关系的思考。在大数据时代，普通的中国人应该如何定位自己的角色？大数据就像一只刚刚长大、尚未关进笼子的富有力量的凶猛野兽，它既能看家护院，又能伤害主人，那么我们应该如何掌控它？

这全在于我们对于自己角色的选择。在大数据时代，我们每个人都有机会成为四种角色，但并不是每个人都有能力作出符合自身最佳利益的选择。

- 不知情者

他们生活在懵懂当中，对于大数据所引发的一系列变化都毫不知

情。他们无知，但又单纯、天真，成为数据收集的第一目标。但与此同时，他们也是超脱的，在不知情的状态中成为一名“幸福的被害者”。

● 知情者

他们了解这个世界正在发生什么，就像他们喜欢一些与大数据有关的话题和书籍。在个人生活中，他们也知道自己成了数据收集方的目标，而且也正在成为这样的数据提供者。因此，他们的内心十分不安，可是又无能为力。

● 参与者

大数据产业的参与者或研究者，他们懂得如何才能保护自己，也知道怎样才可以避免被收集到个人隐私。不过，在这类人的眼中，世界总是黑暗的，他们对未来感到悲观，对技术的进步充满警惕。

● 掌控者

这类人是金字塔的顶端，他们掌握了庞大的数据资源，是大数据时代的精英，既能保护自己，又能成为一名高明的数据收集者，从中获取利益。这些人至少不会是大数据时代的受害者，同时他们又决定了这个时代的发展方向。是魔鬼还是天使，必须由他们自己做出选择。

虽然后三种角色的人可能不会感到快乐，但我们都要力争成为这样的人，而不是稀里糊涂的“不知情者”——他们注定会被这个时代抛弃，被变革的大潮冲击到一个最不起眼的角落。大数据不会等你成熟起来，而是会毫不留情地把你推到一边，然后扬长而去。

对我们来说，大数据是一个新的金矿，是一次新的机遇，尽管它也意味着风险，但它更多的是巨大收益。

如何正确看待它而不走极端呢？

对它即将在我们的生活中产生的影响，我们既不要夸大，也不要低估。如果它在现阶段对你是有害的，那么，小心地远离它；如果是有利的，那么，请谨慎地拥抱它，成为大数据的主人，并且成功地主导它在我们生活中的影响，让它成为你人生新的起点！

附录 打开大数据之门

☆要点

● 如何理解大数据

● 大数据寻宝图

● 思维与行动准备

大数据从2011年开始在世界范围内声名鹊起，2013年是中国的大数据元年。中国人迅速接受了大数据的思维洗礼，从政府到民间层面，都开始推广大数据，使其发挥更大价值。

如果说您已经通过许多同类书籍知道了大数据是什么与可以做什么，那么本书的附录部分则更注重为您解惑和提供实用指导：面对大数据，我们到底应该怎么做？

大数据究竟是什么“数据”？

大数据与商业智能有什么样的区别？

大数据的市场究竟有多大？

我们应该重点发展什么，才能实现超越和后发制人？

我们的优势和劣势在哪里呢？

在各行各业的专家、评论员与参与者的一片喧闹中，我们为您奉上这本书，不期待能有灌顶的功效，却可让您暂时从诸如大数据产业园、大数据日、大数据专委会、大数据专业、大数据实验室或层出不穷的各种大数据峰会接受嘈杂信息的疲劳中摆脱出来，抓住重点，掌握关键，看一看，想一想，为自己找到一个明确和清晰的方向。

☆基本概念——记住4个V

Volume——体量大；

Velocity——快速化；

Variety——类型杂；

Value——价值大。

☆大数据到底有多“大”？

有一家名为IDG的公司对于每年创建和复制的信息体量做过一个计算：

在2011年，大约为1.8ZB；

在2012年，达到了2.8ZB。

根据它的推算，等时间走到2020年时，这个数字将约为40ZB。

当然，也有其他的公司不同意IDG公司的数字，它们预测说道：到2016年时，数据的总量也不过是达到1.3ZB。不过，谷歌公司的统计可能更为震憾——从人类文明开始，一直到2003年，在几十万年间，人类

一共产生了5EB的数据，但到了2010年，产生数据的速度已经到了每两天5EB的程度了。

这表明，在今天的世界，数据不但已经非常之“大”，而且产生得非常之“快”，远非古人甚至十年前的人可以想象。

看到这里，有志于从事大数据产业或投资数据存储的人可能更加富有信心。不过，对普通人而言，我们能知道的无非是另一个关键问题：我们的个人数据是怎么样的？显然，我们为数据总量做着巨大的贡献，并且也享受着这个总量的质变带来的福祉。

不过，不管数据的总量和速度如何变化，我们都要为它设定一个量化标准。在设定了量化标准后，我们就能有一个简单明晰的数值（无论是不是精确）来指导自己或企业对于大数据的判断。这既是必要的一步，也是明智之举。

☆寻宝图——如果你是大数据创业者，请看这里！

作为创业者和技术人员，如果你已对大数据有较深入的了解，你就有必要知道哪些行业才会拥有大数据，即我们将精力投入哪一部分，才能拥有大数据的春天。

在产业链条中，大数据通常分为四类：

科研大数据

科研数据比较古老，实际上在大数据产生前就已经存在了。它们存在于某一些设备、研究资料或者某一些封闭的系统中，拥有科研数据的都是传统的科研机构。它们属于典型的“高富帅”，往往会忽略大众市场。当然，科研大数据的进入门槛也是很高的，往往由国家或大企业主导，个人难以进入。

互联网大数据

互联网大数据肯定是目前的主流，特别是与社交媒体有关的大数据，被认为是大数据产业的爆发点。几乎所有的大数据技术都起源于互联网企业。我们当然也知道它们如雷贯耳的大名，比如百度、谷歌、脸书、雅虎、亚马逊和阿里巴巴。这一行业的驱动力基于两点：一是互联网企业的价值与用户数的平方成正比，也就是“梅特卡夫定律”；二是脸书创始人扎克伯格曾经引用的信息分享理论，即一个人分享的信息每一到两年就会翻一番。

在互联网大数据的产业链中，大型企业占据着绝对的主导地位，它们不仅自身收集和拥有大体量的数据，而且还有平台带动作用，比如阿里巴巴的数据交换平台，360的大型数据中心，百度公司的大数据实验室。所以，中型企业只有开启服务模式才能生存，投入主要精力在外围开发、优化和运作，并同时发展自己的特色，比如豆瓣的“推荐”。

小型公司则属于更低一级的模式，它们情况特殊，虽然拥有一定量的数据，但没有大数据能力，这就催生了一些大数据技术和服务的机会。比如，它们可以为电商网站做个性化推荐和营销分析。还有一些各类广告联盟、移动应用服务平台和提供统计分析、营销服务的公司等，都属于这种情况。

企业大数据

企业的数据比起十几年前并没有数量级的提升，但是在传统基础上加入了非结构化的数据内容。而且，企业大数据与感知大数据有些方面是重叠的，比如企业会部署物联网来收集感知数据。

感知大数据

企业数据是由人来产生的，感知数据是物、传感器、标识等机器产生的。相比之下，感知数据的体量要大得多。有一家公司向我们预测，认为感知数据的总量在2015年将超过社交媒体，并且会达到后者的10～20倍。

我们之所以可以将企业大数据与感知大数据连为一体，划为一个重叠且具备相同性质的产业链条，是因为这两者都涉及传统产业，从经济总量上要比互联网产业大很多。而且重要的是，传统产业自身的大数据能力有限，所以这也是大数据技术和服务企业的主要目标市场，是中小投资者的重要机遇。

从具体的行业讲，对大数据拥有巨大需求的主要集中在公共管理和服务、电信、金融、医疗、零售等方面。不过，在市场竞争激烈的情况下，越是需求巨大的客户，就越难以提供给你轻松进入的黄金机遇，哪怕你的大数据实力是相当优异的。因为不管你走到哪儿，都会发现那些巨头的身影。

☆思维与行动的准备——决策者的板块

作为一个业务决策者，你应具有的大数据观是什么形态的？面临着如此体量巨大的数据，你在思维和行动上要做什么样的准备？

在大数据时代，我们需要新的世界观。大数据已经在技术上为我们开启了一个全新的世界，那么我们就必须主动求变，在思维与行动上对这个世界体现出更新的认知，并高效地转化为结果。

对决策者来说，大数据其实是一种思维，也是战略层面的东西。决策者应该从中看到用户和应用，而不仅是一种技术。但是很明显，许多企业的决策者都在这方面缺位了，他们醉心于技术层面的演进，缺乏宏观思维和布局。

旧的认知——数据是一种稀缺资源。

这种认知直接导致了决策者的小农心态，不去关注数据测量和海量

的数据收集，而是总幻想可以从最少的数据中挤压出最多的信息。

新的认知——大数据的关键在于“大”。

决策者自己要有勇气参与大数据的游戏并且取得胜利，为自己树立“大”的概念，去收集全数据，而不是习惯于过去的抽样处理和分析。决策必须建立在全数据的基础上，全面和客观地去分析所有因素，并将此作为自己的一种责任和信仰。

决策者需要具备的大数据观也很简单：对我们来说数据不是累赘，而是财富；数据不管用过没有，都要保存下来，从而逐渐将“成本”转化为“利润”。而且，必须尽量地减少自己的主观性。

1.让数据收集工具决定收集哪些信息，去哪里收集。

2.如果我们的分析过程带有天然的主观性，比如民意调查或街头采访等，那么在做出数据采集的决策前，你有责任为它设计更客观的前提，比如通过设置很多问题来减少主观误差。

3.你要尽可能地把数据采集和存储纳入一个共享的平台，也就是建立一个基础框架，而不是来一个业务就做一种不同的采集和存储方案。并且，你还需要在数据采集的过程中引入激励机制，为决策做最充足的准备，收集最丰富的信息。